Distintas perspectivas a investigación en pregrado

DCSeminario #2

Fernando Florenzano Hernández
faflorenzano@uc.cl

Contenidos

Frases típicas

Visualizador RDF

Algoritmos eficientes en EI

Conclusiones

Contenidos

Frases típicas

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

“¿En que aportaría? **Está todo resuelto**”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

“¿En que aportaría? **Está todo resuelto**”

“¿Qué gano yo? **Son solo créditos**”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

“¿En que aportaría? **Está todo resuelto**”

“¿Qué gano yo? **Son solo créditos**”

“¿Y si **no resulta**?”

Frases típicas

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

“¿En que aportaría? **Está todo resuelto**”

“¿Qué gano yo? **Son solo créditos**”

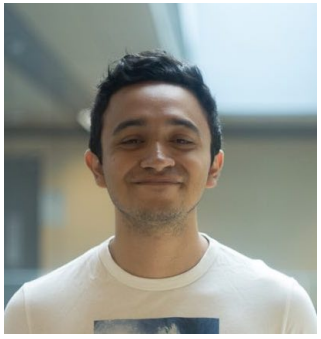
“¿Y si **no resulta**?”

“Es terrible **fome** investigar”

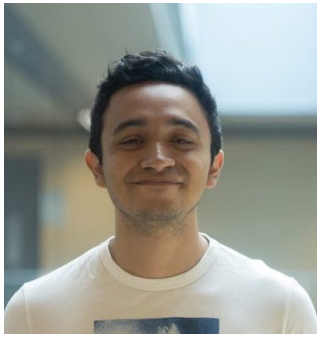
Contenidos

Erase una vez...

Erase una vez...

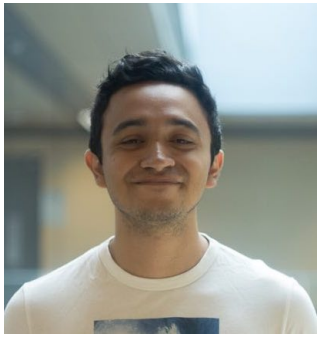


Erase una vez...



...un **joven** Fernando, que no sabía que *major* tomar.

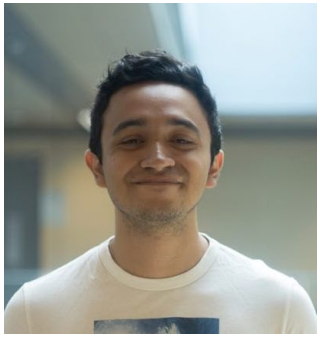
Erase una vez...



...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

Erase una vez...

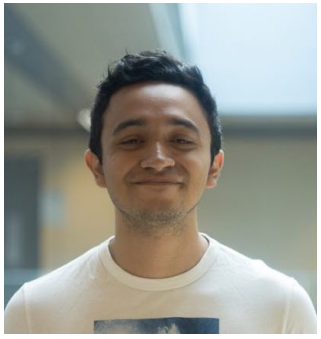


...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

- **Programación Avanzada**

Erase una vez...

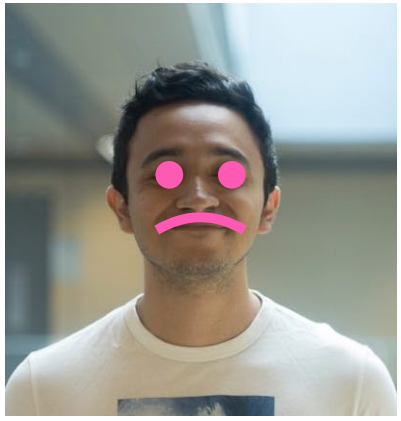


...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

- **Programación Avanzada**
- **Matemáticas Discretas**

Erase una vez...

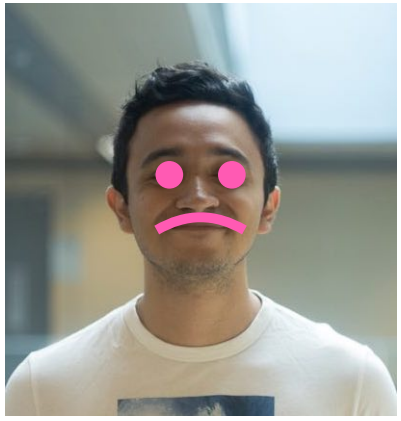


...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

- **Programación Avanzada**
- **Matemáticas Discretas**

Erase una vez...



...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

- **Programación Avanzada**
- ~~**Matemáticas Discretas**~~

Erase una vez...

...un **joven** Fernando, que no sabía que *major* tomar.

Computación parecía interesante, así que tomó los cursos:

- **Programación Avanzada**
- ~~**Matemáticas Discretas**~~

Tomó vuelo...

Tomó vuelo...

Al siguente semestre, continuó tomando cursos de computación:

Tomó vuelo...

Al siguente semestre, continuó tomando cursos de computación:

Juan Reutter - Bases de Datos

Tomó vuelo...

Al siguente semestre, continuó tomando cursos de computación:

Juan Reutter - Bases de Datos

Cristian Riveros - Matemáticas Discretas

Se acercó a hablar

Se acercó a hablar

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena,

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena, ¿y en qué?

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena, ¿y en qué?

Fernando: ...

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena, ¿y en qué?

Fernando: ... ¿bases de datos?

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena, ¿y en qué?

Fernando: … ¿bases de datos?

Juan: …

Se acercó a hablar

Fernando: Profe, me gustaría trabajar en una investigación.

Juan: Buena, ¿y en qué?

Fernando: ... ¿bases de datos?

Juan: ...

Juan: Ya mira, tengo estos temas...

RDF

RDF

RDF es un modelo de datos para metadatos. Se basa en expresar todo como **triples** de objetos que se relacionan.

RDF

RDF es un modelo de datos para metadatos. Se basa en expresar todo como **triples** de objetos que se relacionan.

```
:K_Bacon     :Acts_in   :Crazy_Stupid_Love.
```

RDF

RDF es un modelo de datos para metadatos. Se basa en expresar todo como **triples** de objetos que se relacionan.

```
:K_Bacon     :Acts_in   :Crazy_Stupid_Love.
:R_Gosling   :Acts_in   :Crazy_Stupid_Love.
```

RDF

RDF es un modelo de datos para metadatos. Se basa en expresar todo como **triples** de objetos que se relacionan.

```
:K_Bacon     :Acts_in   :Crazy_Stupid_Love.
:R_Gosling   :Acts_in   :Crazy_Stupid_Love.
:J_Moore     :Acts_in   :Crazy_Stupid_Love.
```

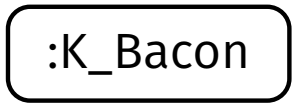
RDF

RDF es un modelo de datos para metadatos. Se basa en expresar todo como **triples** de objetos que se relacionan.

```
:K_Bacon     :Acts_in   :Crazy_Stupid_Love.
:R_Gosling   :Acts_in   :Crazy_Stupid_Love.
:J_Moore     :Acts_in   :Crazy_Stupid_Love.
:K_Bacon     :Directs   :Loverboy.
```

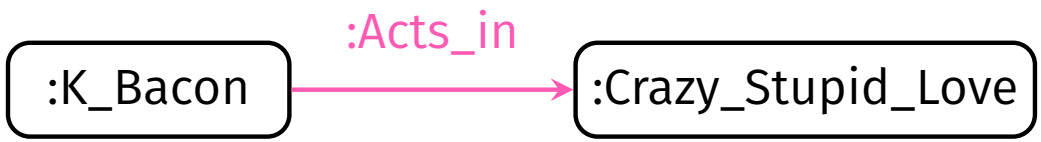
RDF como grafo

RDF como grafo

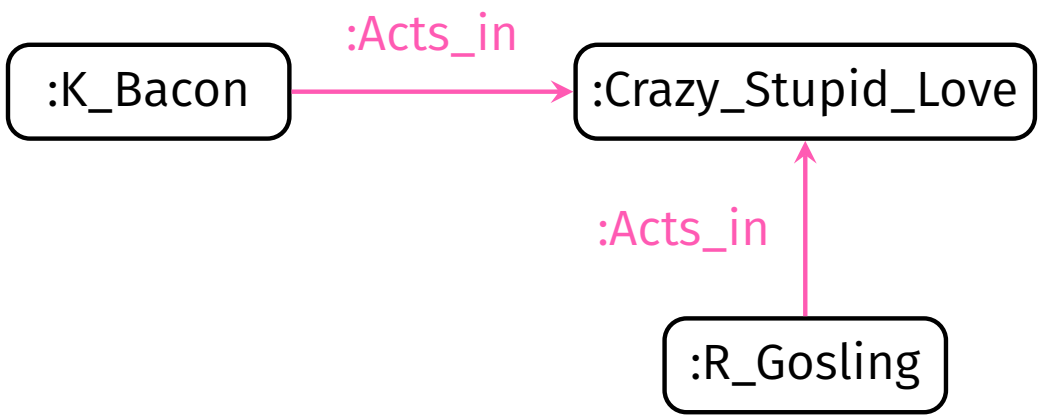


RDF como grafo

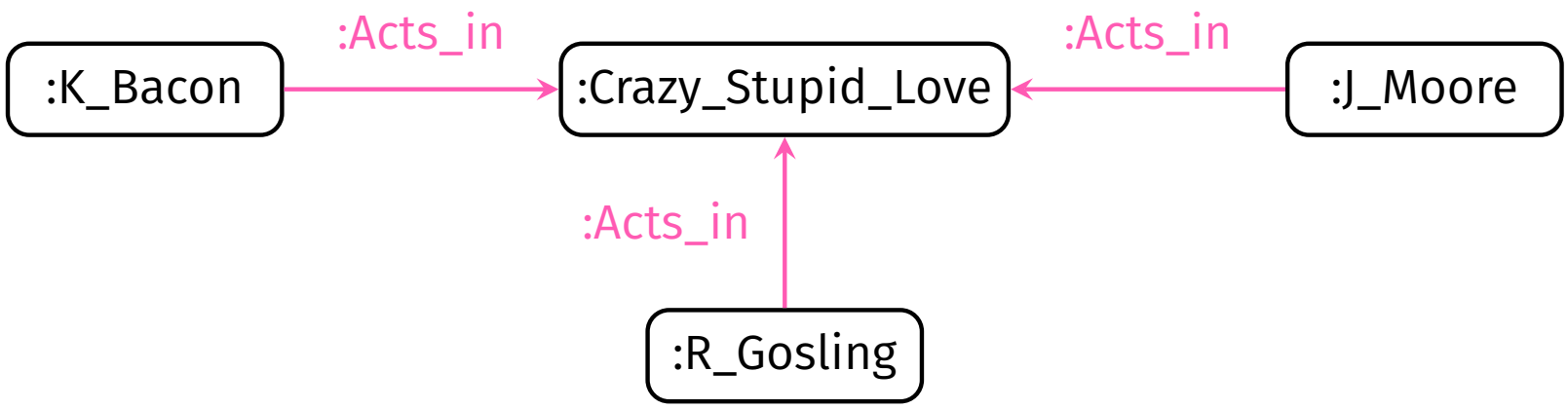
RDF como grafo



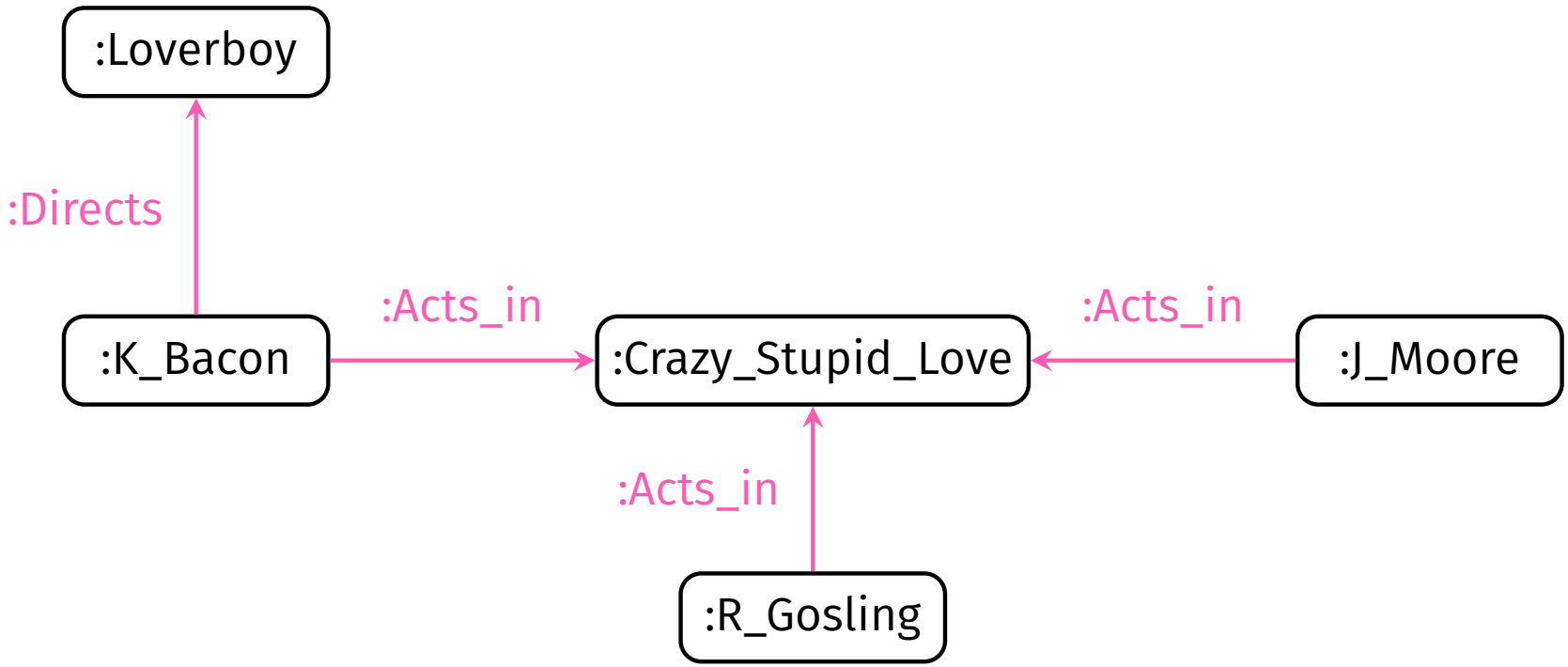
RDF como grafo



RDF como grafo



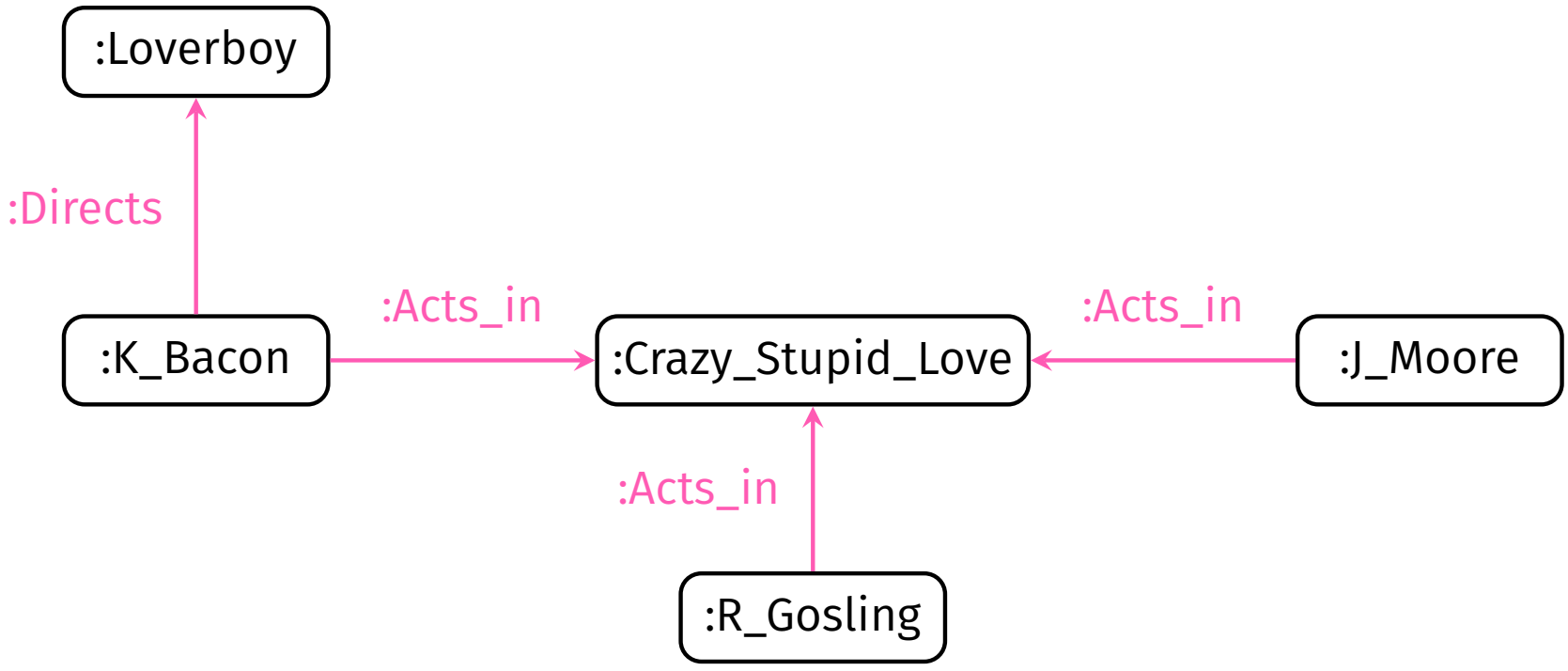
RDF como grafo



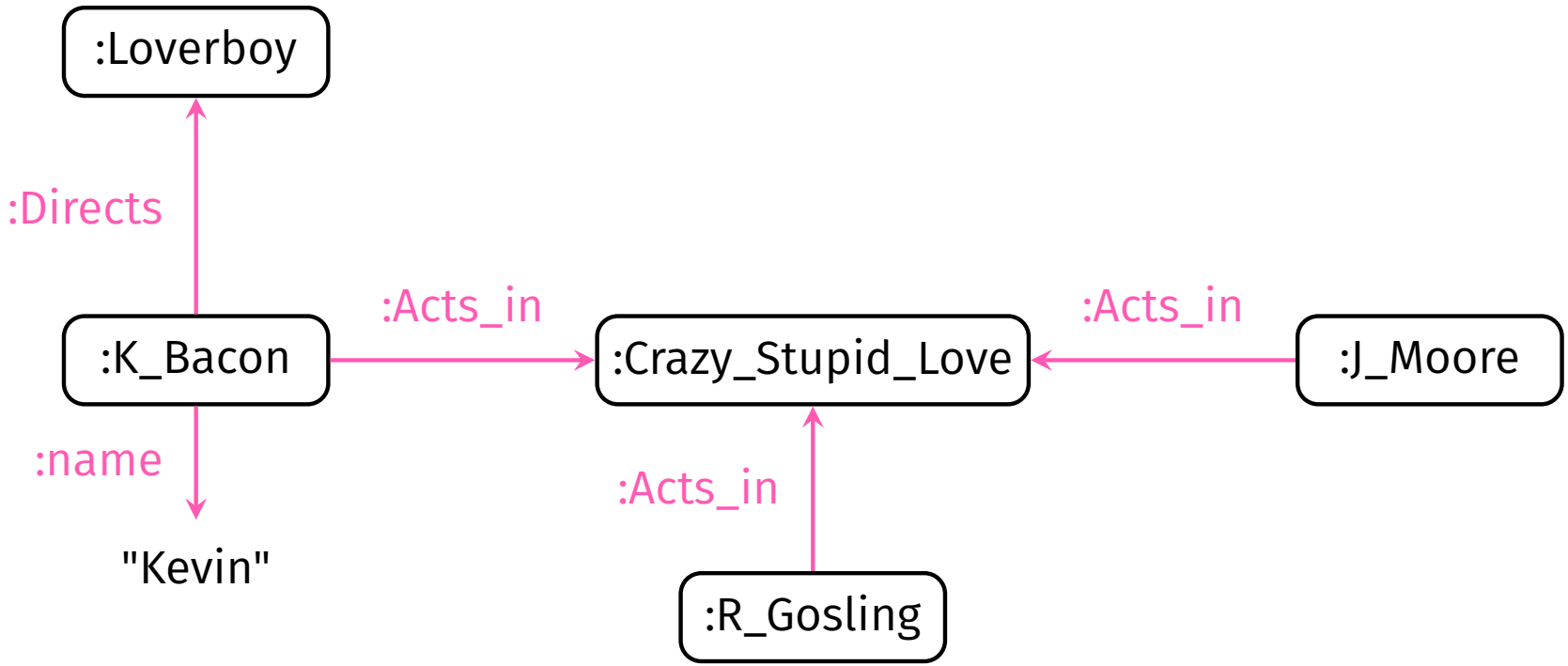
RDF puede agregar propiedades

```
:K_Bacon              :name    "Kevin".
:R_Gosling            :name    "Ryan".
:J_Moore              :name    "Julianne".
:Crazy_Stupid_Love    :title   "Crazy Stupid Love".
:Loverboy             :title   "Loverboy".
```

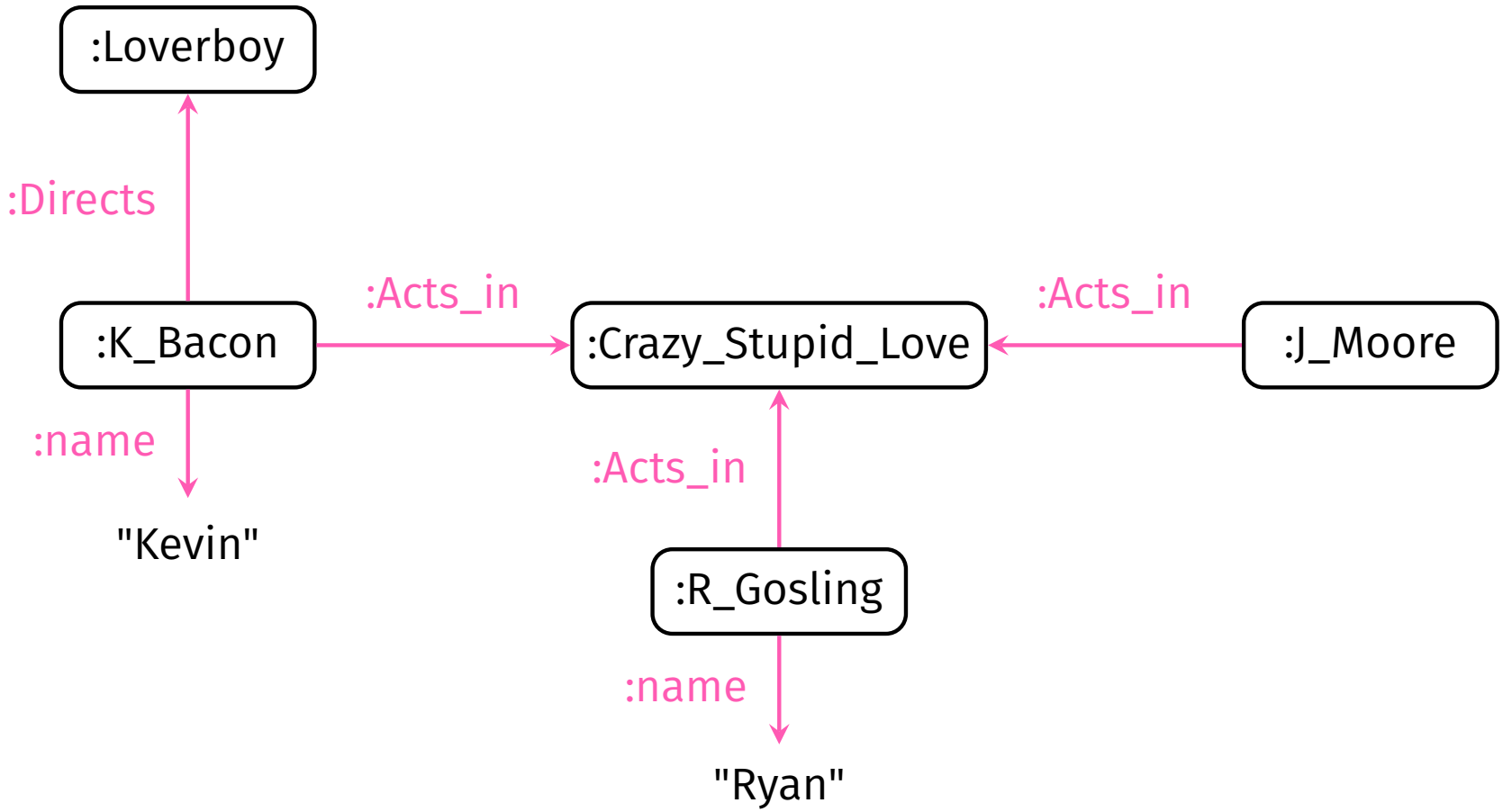
Recursos con propiedades



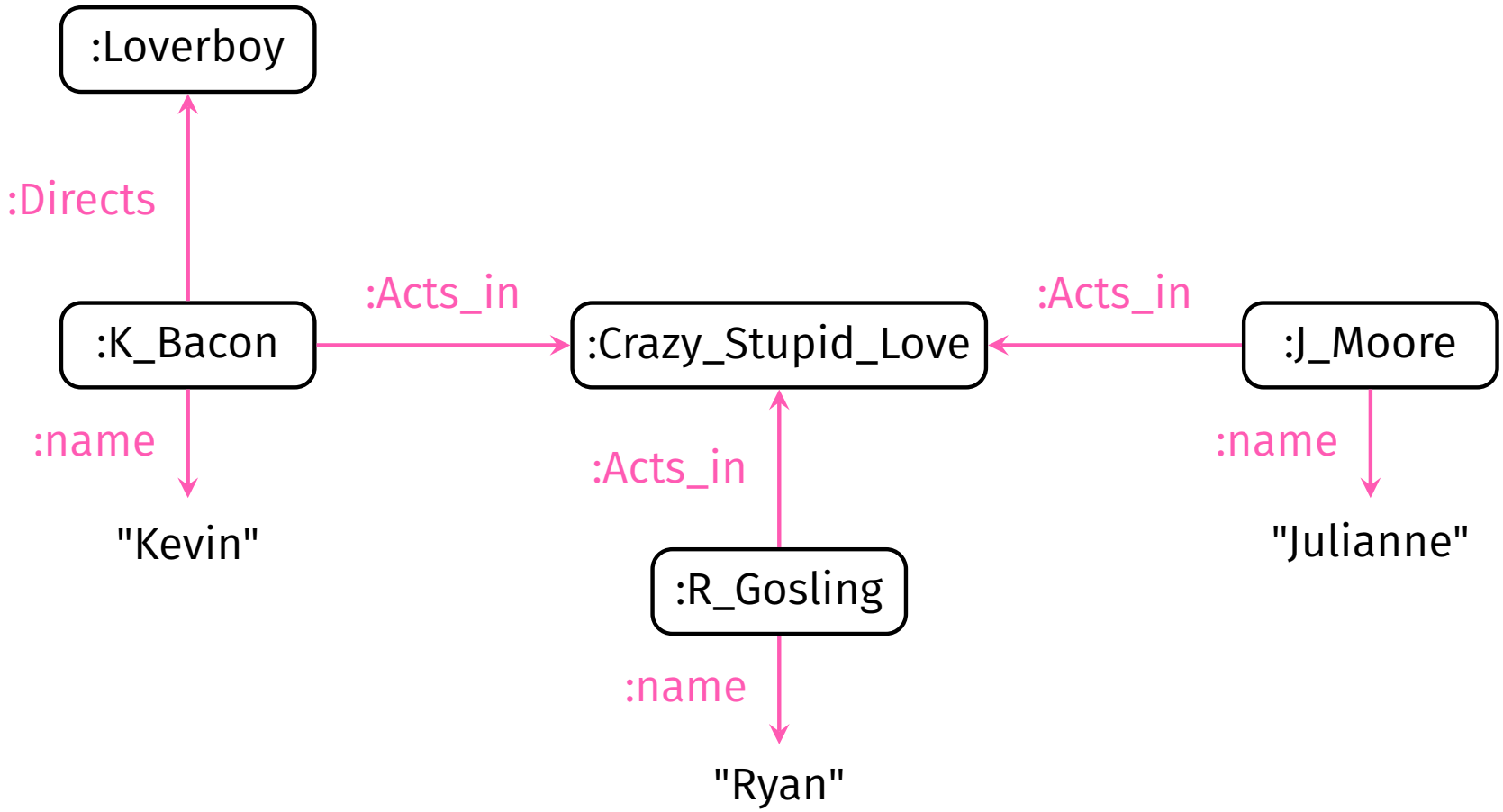
Recursos con propiedades



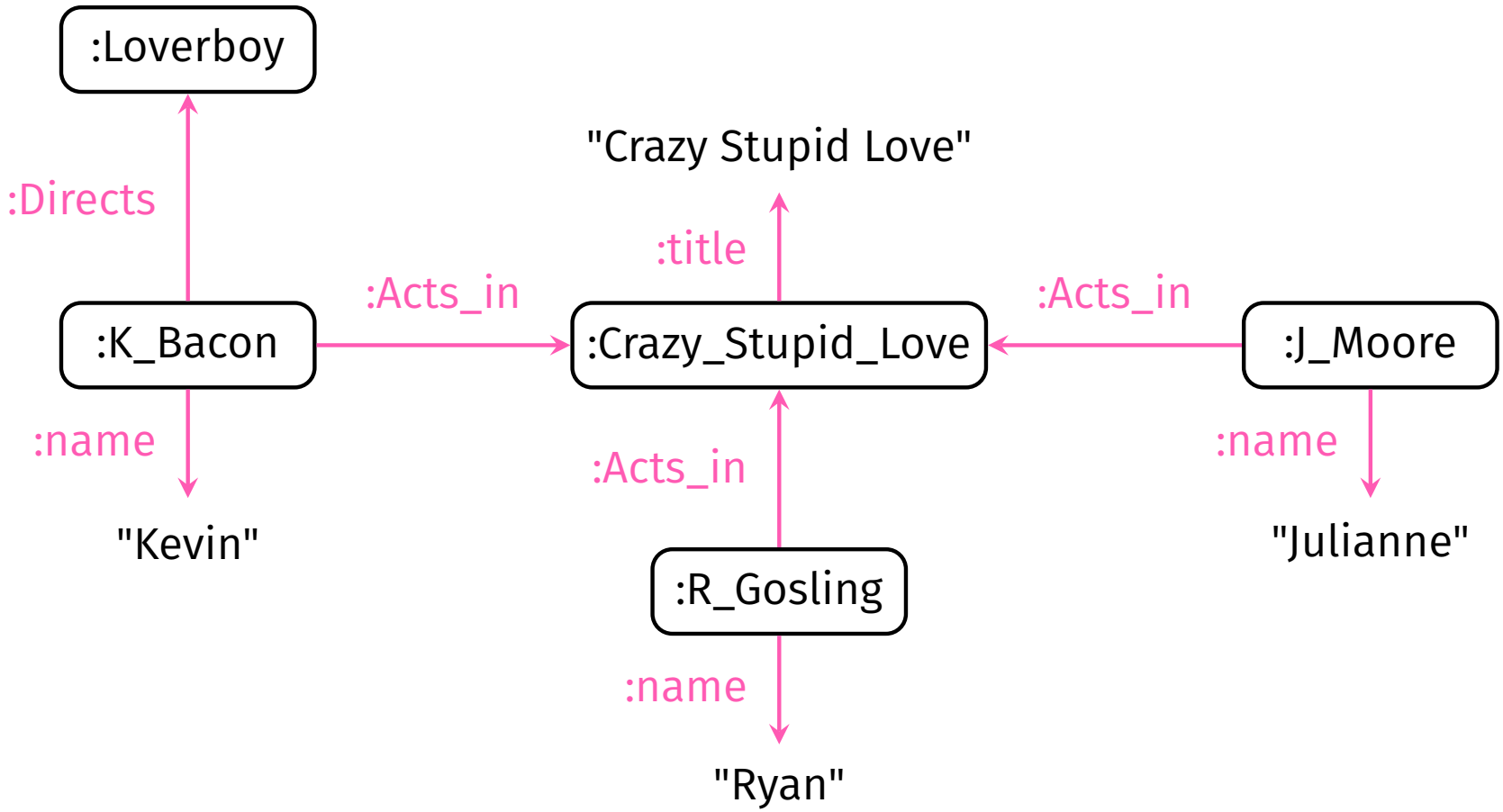
Recursos con propiedades



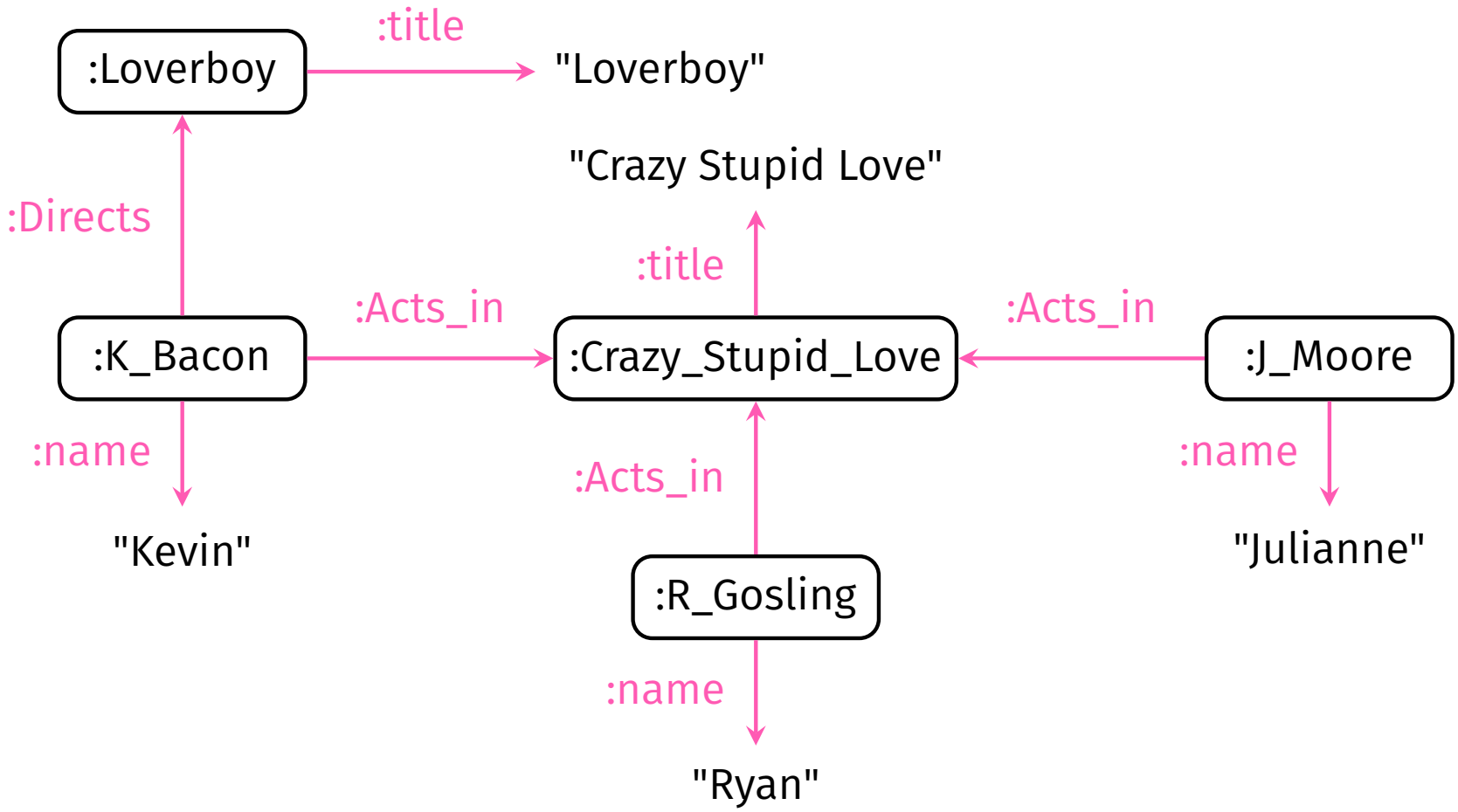
Recursos con propiedades



Recursos con propiedades



Recursos con propiedades



SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.

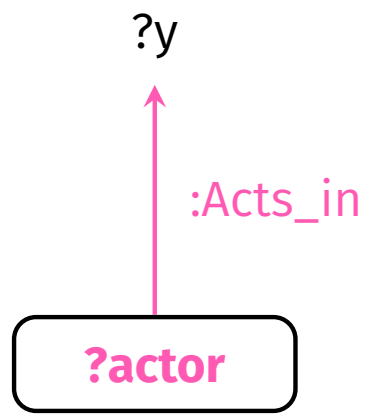
SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.

?actor

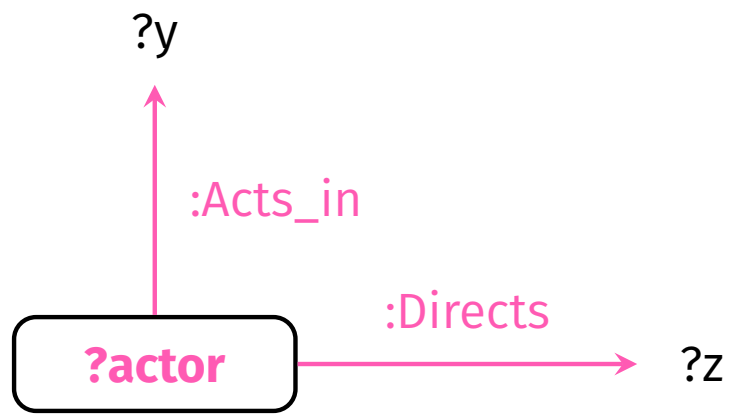
SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.



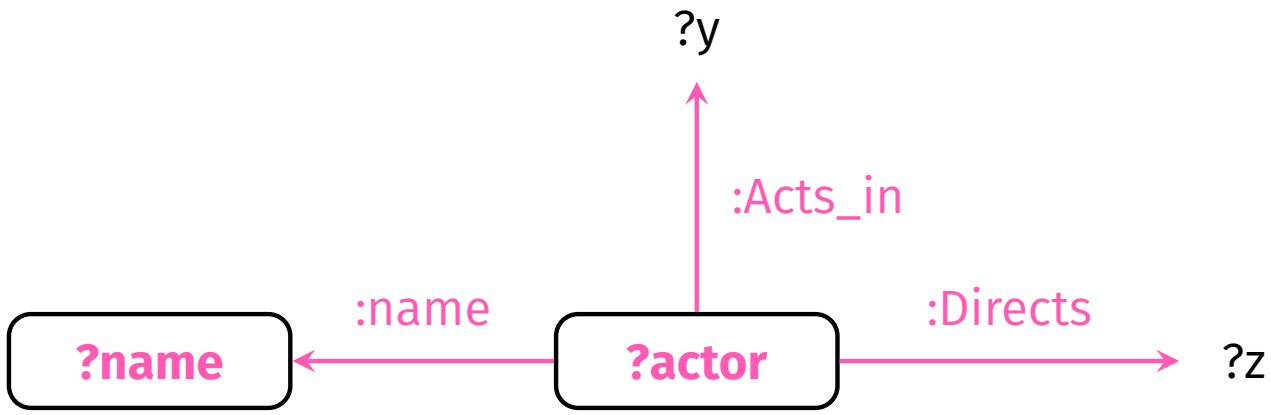
SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.



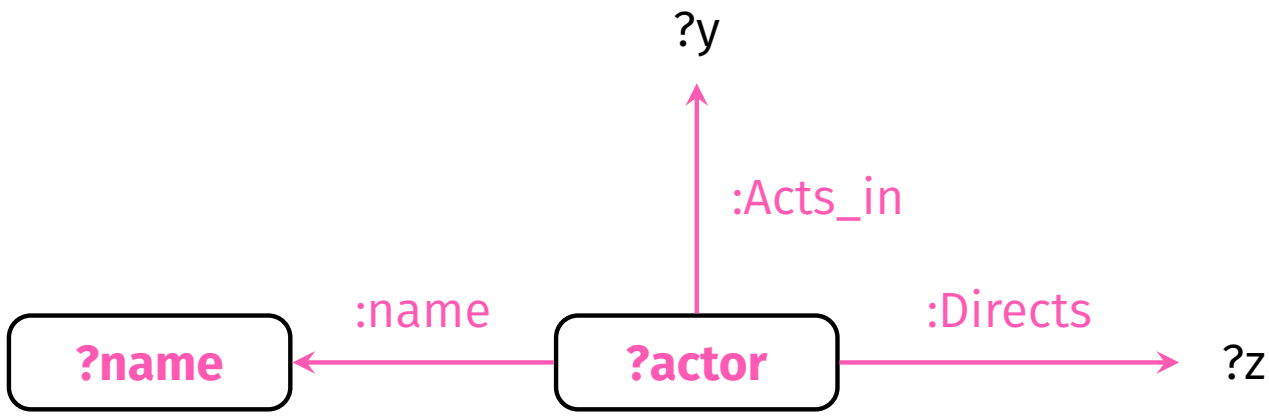
SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.



SPARQL: consultas semánticas

SPARQL es el lenguaje de consulta para **RDF**. Un usuario accede a un punto de entrada de una base de datos RDF, envía una consulta y recibe una respuesta.



actor	name
:K_Bacon	"Kevin"

Problema en RDF

Problema en RDF

¿Y si no sé que hay en una base de datos RDF?
¿Cómo la consulto?

Problema en RDF

Problema en RDF

- Para realizar una consulta **con sentido**, necesito saber **previamente** que tipo de objetos o relaciones hay en la base de datos RDF.

Problema en RDF

- Para realizar una consulta **con sentido**, necesito saber **previamente** que tipo de objetos o relaciones hay en la base de datos RDF.
- Como esto se procesa a través de la web, la cantidad de respuestas se limita para que no sea **tan grande**.

Problema en RDF

- Para realizar una consulta **con sentido**, necesito saber **previamente** que tipo de objetos o relaciones hay en la base de datos RDF.
- Como esto se procesa a través de la web, la cantidad de respuestas se limita para que no sea **tan grande**.
- Como todo se almacena en forma de triples, **no hay fácil forma** de saber la semántica de los datos sin consultarlos antes.

Problema en RDF

- Para realizar una consulta **con sentido**, necesito saber **previamente** que tipo de objetos o relaciones hay en la base de datos RDF.
- Como esto se procesa a través de la web, la cantidad de respuestas se limita para que no sea **tan grande**.
- Como todo se almacena en forma de triples, **no hay fácil forma** de saber la semántica de los datos sin consultarlos antes.

Uno accede ciegamente a hacer consultas a una base de datos RDF.

Idea

Idea

Crear una interfaz gráfica que me permita entender el contenido de una base de datos **RDF** de forma **visual**.

Idea

Crear una interfaz gráfica que me permita entender el contenido de una base de datos **RDF** de forma **visual**.

- Entender que **tipo** de objetos existen en la base de datos.

Idea

Crear una interfaz gráfica que me permita entender el contenido de una base de datos **RDF** de forma **visual**.

- Entender que **tipo** de objetos existen en la base de datos.
- Entender como se **relacionan** estos tipos de objetos.

Desarrollo

Desarrollo

Mis tareas:

- Aprender a **procesar** datos para luego visualizarlos.

Desarrollo

Mis tareas:

- Aprender a **procesar** datos para luego visualizarlos.
- Aprender **JavaScript** y **D3**.

Desarrollo

Mis tareas:

- Aprender a **procesar** datos para luego visualizarlos.
- Aprender **JavaScript** y **D3**.
- Aprender a montar un **servidor web**.

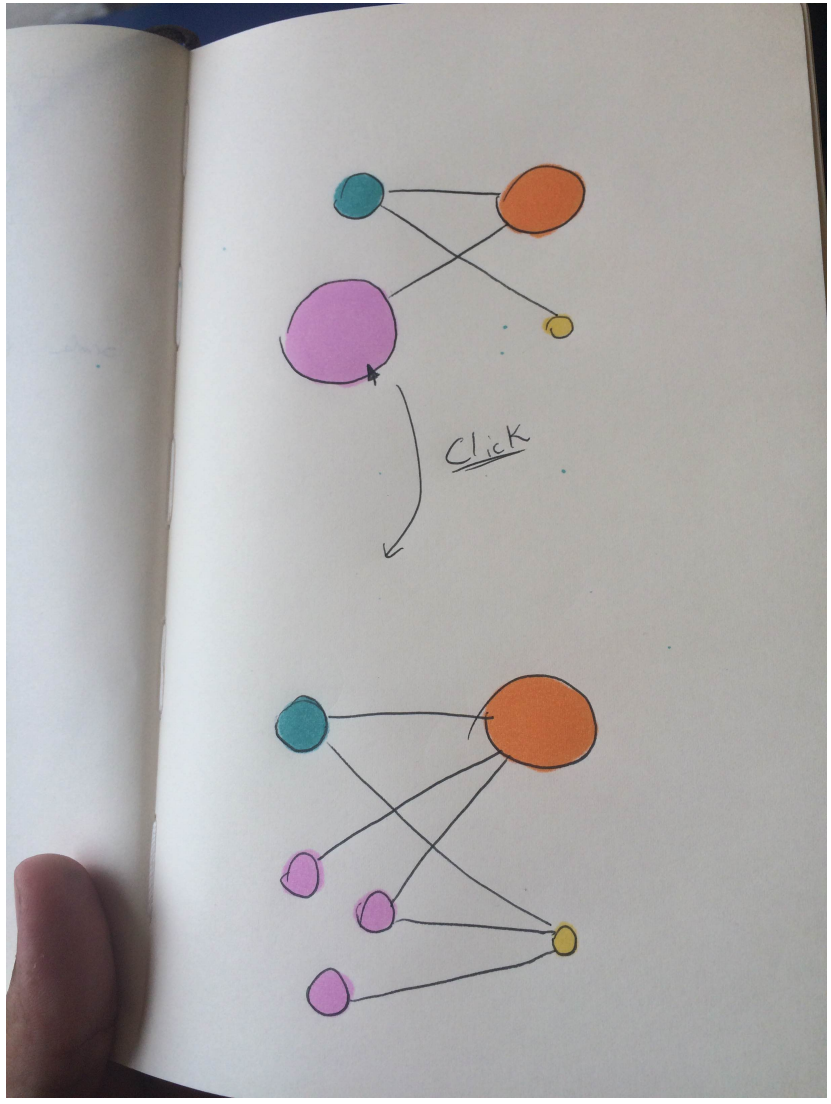
Desarrollo

Mis tareas:

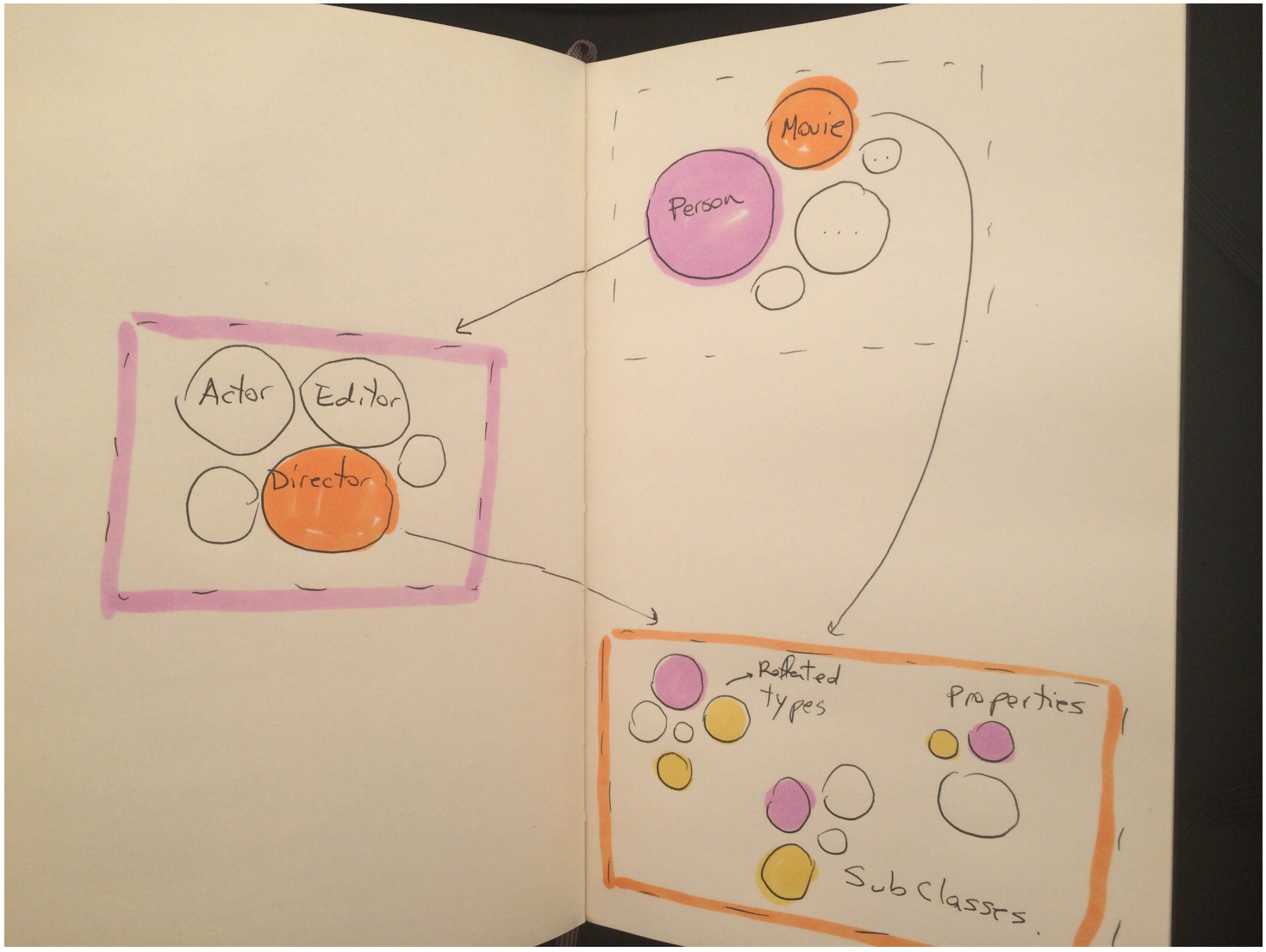
- Aprender a **procesar** datos para luego visualizarlos.
- Aprender **JavaScript** y **D3**.
- Aprender a montar un **servidor web**.

Cosas que **no** aprendí en el curso **Bases de Datos**, pero aprendí trabajando en esta idea.

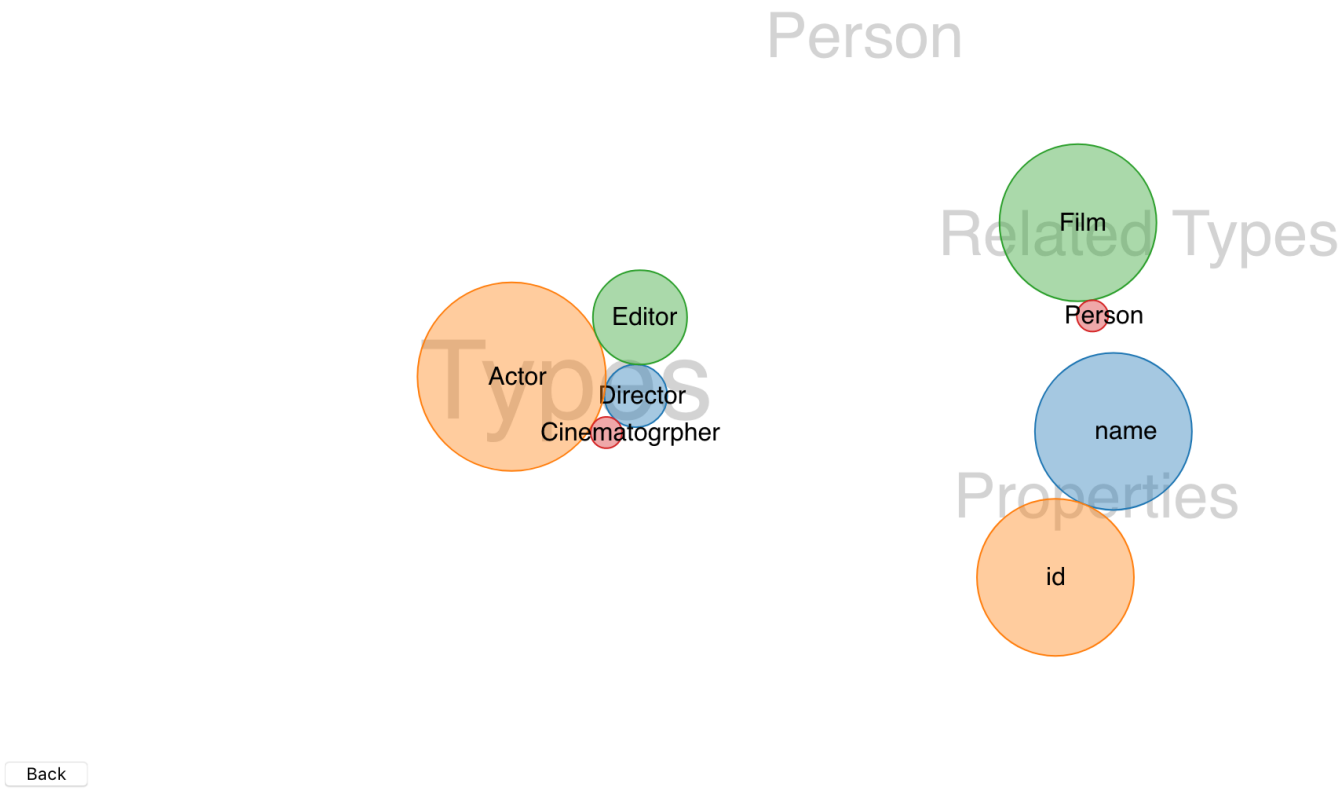
Desarrollo



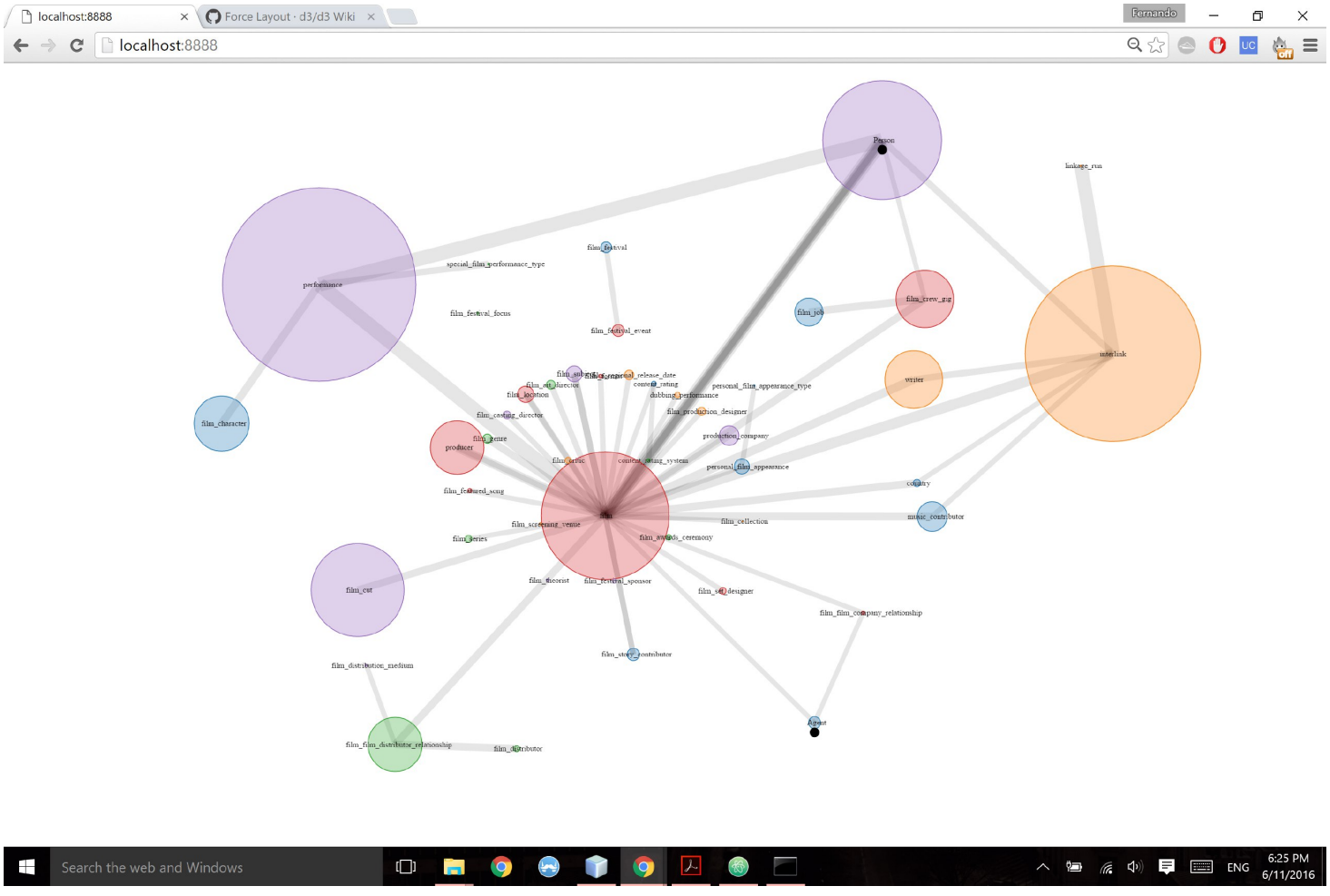
Desarrollo


Movie
Person
Actor
Editor
Director
Related types
Properties
Subclasses.

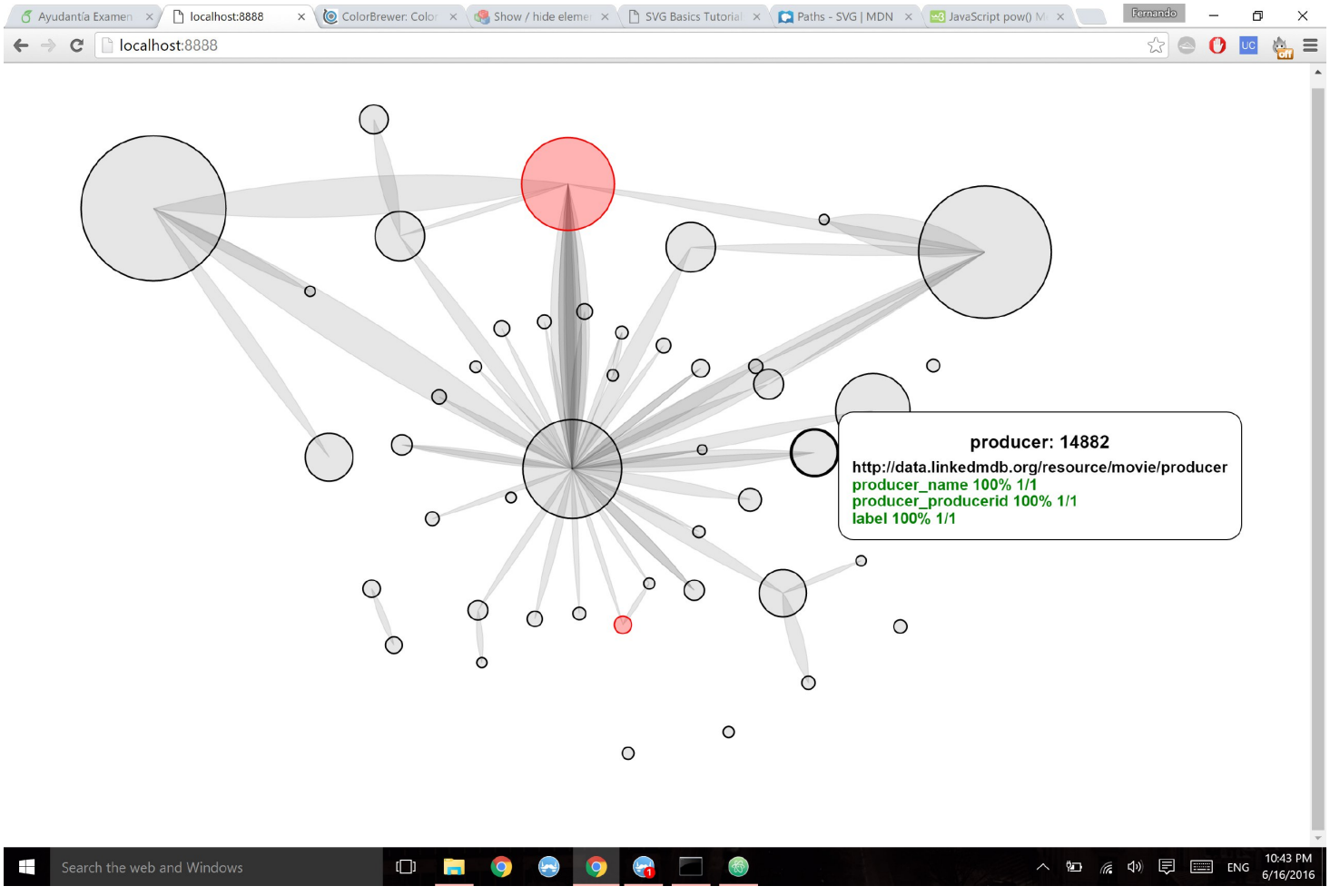
Desarrollo


Person
Types
Editor
Actor
Director
Cinematogrpher
Related Types
Film
Person
Properties
name
id
Back

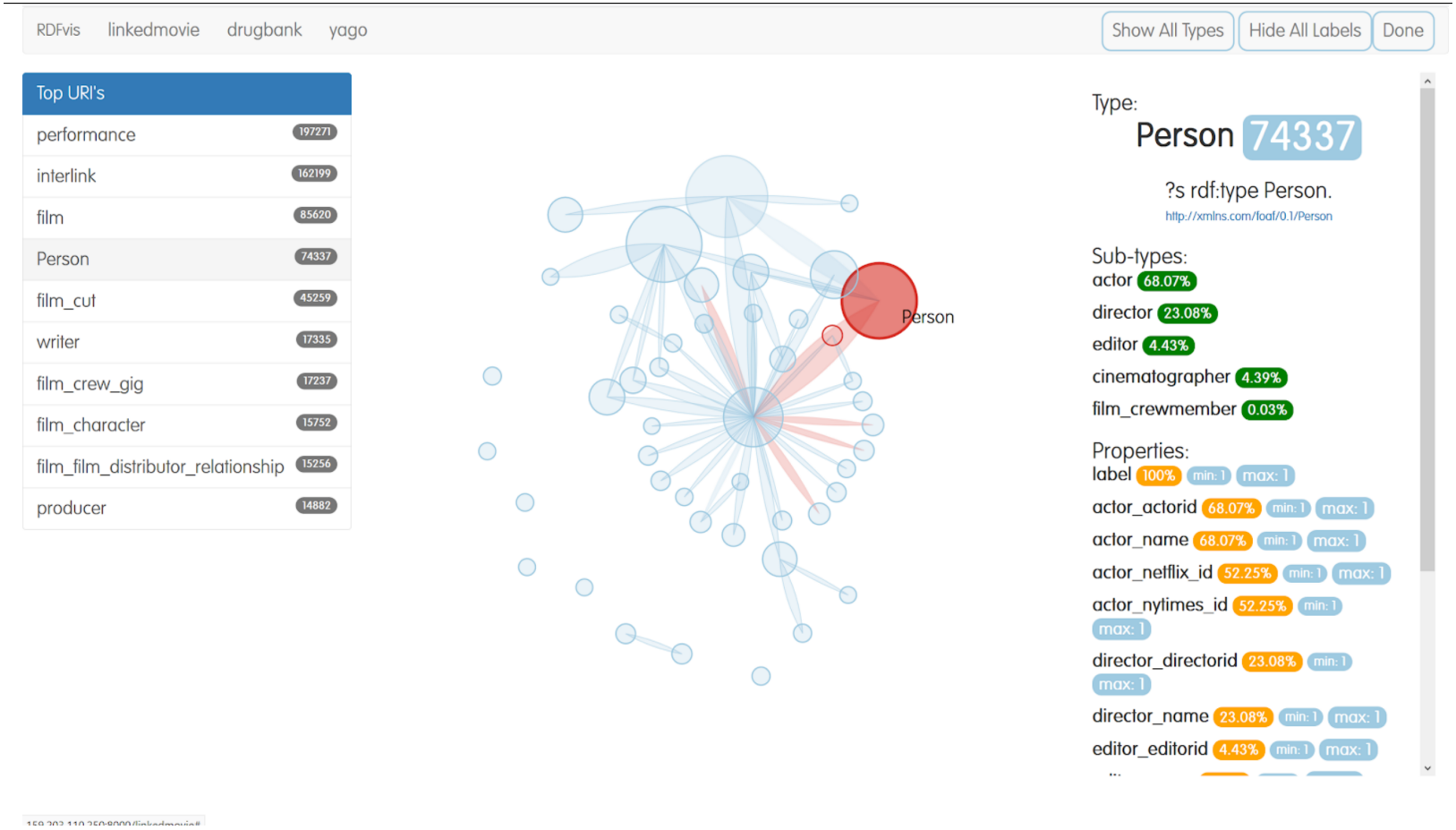
Desarrollo


localhost:8888
Force Layout · d3/d3 Wiki
Fernando
localhost:8888
Person
performance
film_character
film_festival
special_film_performance_type
film_festival_focus
film_festival_event
film_job
film_crew_gig
linkage_run
interlink
writer
film_location
film_casting_director
film_production_designer
production_company
personal_film_appearance_type
dubbing_performance
personal_film_appearance
producer
film_genre
film_artisc
content_rating_system
film_featured_song
film_screening_venue
film_series
film_awards_ceremony
film_collection
country
music_contributor
film
film_cast
film_theorist
film_festival_sponsor
film_set_designer
film_film_company_relationship
film_story_contributor
film_distribution_medium
film_film_distributor_relationship
film_distributor
Agent
Search the web and Windows
ENG
6:25 PM
6/11/2016

Desarrollo


Ayudantía Examen
localhost:8888
ColorBrewer: Color
Show / hide eleme
SVG Basics Tutorial
Paths - SVG | MDN
JavaScript pow() M
Fernando
localhost:8888
producer: 14882
http://data.linkedmdb.org/resource/movie/producer
producer_name 100% 1/1
producer_producerid 100% 1/1
label 100% 1/1
Search the web and Windows
ENG
10:43 PM
6/16/2016

Desarrollo


RDFvis
linkedmovie
drugbank
yago
Show All Types
Hide All Labels
Done
Top URI's
performance 197271
interlink 162199
film 85620
Person 74337
film_cut 45259
writer 17335
film_crew_gig 17237
film_character 15752
film_film_distributor_relationship 15256
producer 14882
Person
Type:
Person 74337
?s rdf:type Person.
http://xmlns.com/foaf/0.1/Person
Sub-types:
actor 68.07%
director 23.08%
editor 4.43%
cinematographer 4.39%
film_crewmember 0.03%
Properties:
label 100% min: 1 max: 1
actor_actorid 68.07% min: 1 max: 1
actor_name 68.07% min: 1 max: 1
actor_netflix_id 52.25% min: 1 max: 1
actor_nytimes_id 52.25% min: 1 max: 1
director_directorid 23.08% min: 1 max: 1
director_name 23.08% min: 1 max: 1
editor_editorid 4.43% min: 1 max: 1

Resultados

A visual aide for understanding endpoint data

Fernando Florenzano[1,2], Denis Parra[1], Juan Reutter[1,2], and Freddie Venegas[1]

[1] Pontificia Universidad Católica de Chile
[2] Center for Semantic Web research, CL

Abstract. In order to pose queries on SPARQL endpoints, users need to understand the underlying structure of the data that is stored. Unfortunately, and despite the importance of endpoints in the Semantic Web infrastructure, in most (if not all) publicly available endpoints the only way of understanding this structure is by performing a considerable number of probe queries, perhaps inspired in a few examples that are also made available.

This paper looks into the problem of providing additional information for SPARQL-fluent users that need to query a RDF dataset they are not familiar with. We set up to understand what is the essential information that a user needs to query a SPARQL dataset, and then propose a visualisation that can effectively help users learn this information. This visualisation consists of a labelled graph whose nodes are the different types of entities in the RDF dataset, and where two types are related if entities of these types appear related in the RDF dataset. We illustrate our visualisation using the Linked Movie Database dataset.

Resultados

Resultados

Contenidos

Un tiempo después...

Un tiempo después...

... un Fernando, **un poco menos joven**,

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

A finales de ese semestre Cristian se le acerca.

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

A finales de ese semestre Cristian se le acerca.

Cristian: Fernando, ¿te interesaría trabajar en temas teóricos?

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

A finales de ese semestre Cristian se le acerca.

Cristian: Fernando, ¿te interesaría trabajar en temas teóricos?
Fernando: ...

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

A finales de ese semestre Cristian se le acerca.

Cristian: Fernando, ¿te interesaría trabajar en temas teóricos?
Fernando: ...
Cristian: Buena,

Un tiempo después...

... un Fernando, **un poco menos joven**, venció su miedo hacia **Matemáticas Discretas** y postuló a la ayudantía.

A finales de ese semestre Cristian se le acerca.

Cristian: Fernando, ¿te interesaría trabajar en temas teóricos?
Fernando: ...
Cristian: Buena, tengo estos temas...

Extracción de Información

Extracción de Información

Extracción de Información es un tipo de **recuperación de información** cuyo objetivo es extraer **automáticamente** información **estructurada o semiestructurada** desde documentos **legibles** por una computadora.

Extracción de Información

Extracción de Información es un tipo de **recuperación de información** cuyo objetivo es extraer **automáticamente** información **estructurada o semiestructurada** desde documentos **legibles** por una computadora.

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

Extracción de Información

Extracción de Información es un tipo de **recuperación de información** cuyo objetivo es extraer **automáticamente** información **estructurada o semiestructurada** desde documentos **legibles** por una computadora.

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”

Extracción de Información

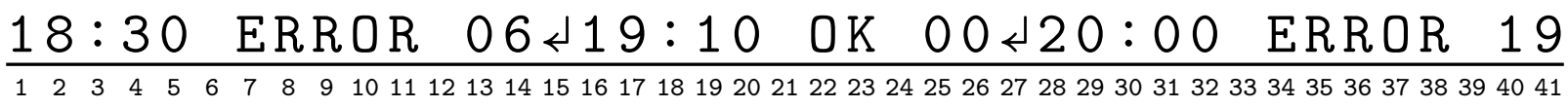
```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”

Extracción de Información

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”



Extracción de Información

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”

```
18:30 ERROR 06↲19:10 OK 00↲20:00 ERROR 19
1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
```

Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

Extracción de Información

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”

```
18:30 ERROR 06↲19:10 OK 00↲20:00 ERROR 19
1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
```

Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

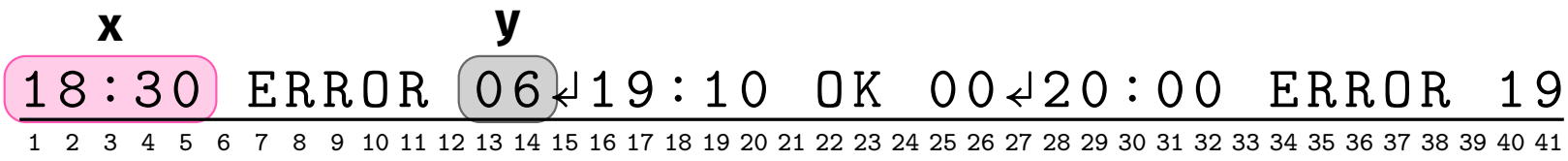
Resultado: intervalos

x	y

Extracción de Información

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”



Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

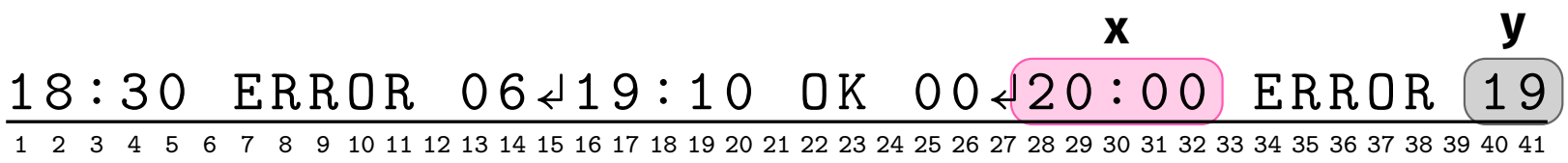
Resultado: intervalos

x	y
$[1, 6\rangle$	$[13, 15\rangle$

Extracción de Información

```
18:30 ERROR 06
19:10 OK 00
20:00 ERROR 19
```

“Extraer todos los pares (tiempo,id) de eventos ERROR”



Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

Resultado: intervalos

x	y
$[1, 6\rangle$	$[13, 15\rangle$
$[28, 33\rangle$	$[40, 42\rangle$

Problema a resolver

Problema: Evaluación de reglas en extracción de información.

```
18:30 ERROR 06↲19:10 OK 00↲20:00 ERROR 19
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
```

Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

Resultado: intervalos

x	y
$[1, 6\rangle$	$[13, 15\rangle$
$[28, 33\rangle$	$[40, 42\rangle$

Problema a resolver

Problema: Evaluación de reglas en extracción de información.
Input: Fórmula RGX *R* y documento *d*.

18:30 ERROR 06↲19:10 OK 00↲20:00 ERROR 19

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

Resultado: intervalos

x	y
$[1, 6\rangle$	$[13, 15\rangle$
$[28, 33\rangle$	$[40, 42\rangle$

Problema a resolver

Problema: Evaluación de reglas en extracción de información.
Input: Fórmula RGX *R* y documento *d*.
Output: **Enumerar** todos los intervalos de *d* que calzan con *R*.

```
18:30 ERROR 06↲19:10 OK 00↲20:00 ERROR 19
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
```

Regla: fórmula RGX

$$\Sigma^* \cdot \mathbf{x}\{\delta\delta : \delta\delta\} \cdot \text{␣ERROR␣} \cdot \mathbf{y}\{\delta\delta\} \cdot \Sigma^*$$

$$\delta = (0 + 1 + \ldots + 9)$$

Resultado: intervalos

x	y
$[1, 6\rangle$	$[13, 15\rangle$
$[28, 33\rangle$	$[40, 42\rangle$

Algoritmos de demora constante

Algoritmos de demora constante

Definición

Dada una regla RGX R y un documento d,
un **algoritmo de demora constante** es un algoritmo de **dos fases** :

Algoritmos de demora constante

Definición

Dada una regla RGX R y un documento d,
un **algoritmo de demora constante** es un algoritmo de **dos fases** :

1. Fase de **preprocesamiento**: de tiempo lineal en $|d|$ y, ojalá, lineal en $|R|$.

Algoritmos de demora constante

Definición

Dada una regla RGX R y un documento d,
un **algoritmo de demora constante** es un algoritmo de **dos fases** :

1. Fase de **preprocesamiento**: de tiempo lineal en $|d|$ y, ojalá, lineal en $|R|$.
2. Fase de **enumeración**: de tiempo **constante** entre respuestas consecutivas.

Algoritmos de demora constante

Definición

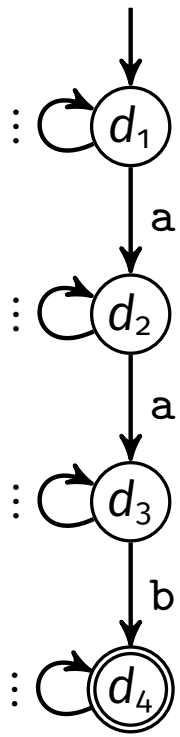
Dada una regla RGX R y un documento d,
un **algoritmo de demora constante** es un algoritmo de **dos fases** :

1. Fase de **preprocesamiento**: de tiempo lineal en $|d|$ y, ojalá, lineal en $|R|$.
2. Fase de **enumeración**: de tiempo **constante** entre respuestas consecutivas.

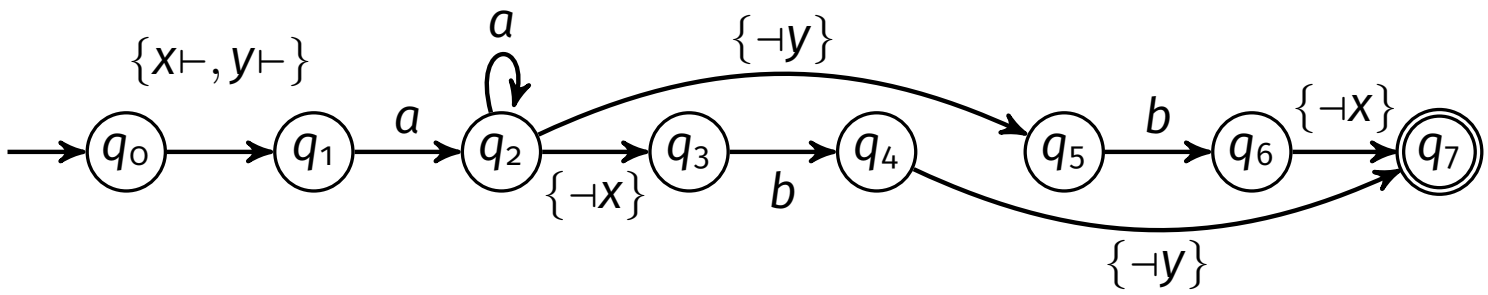
¿Es posible encontrar un algoritmo de demora constante **eficiente** para fórmulas RGX?

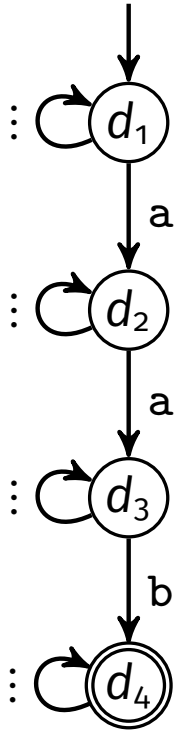
Idea

Idea

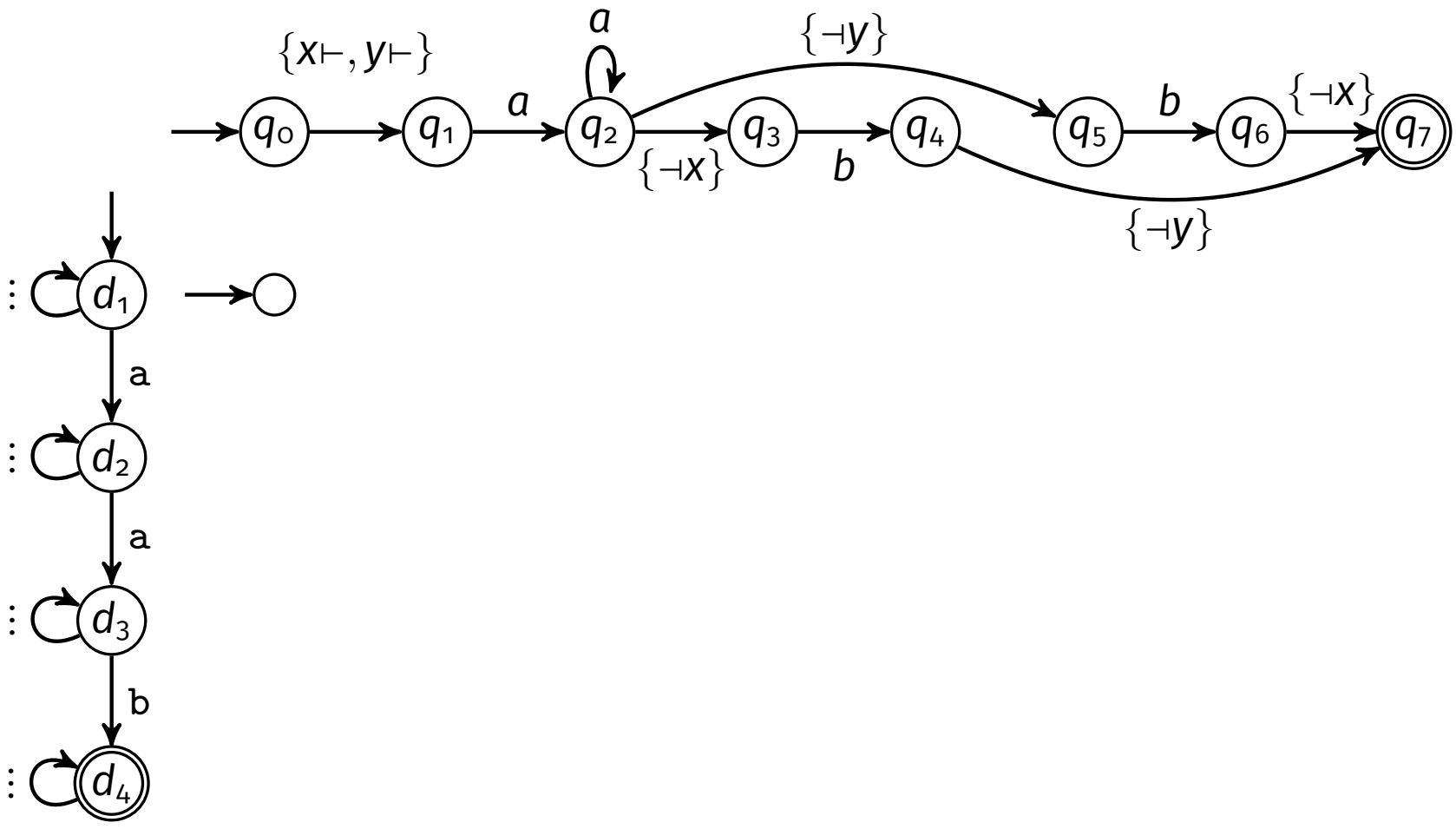


Idea

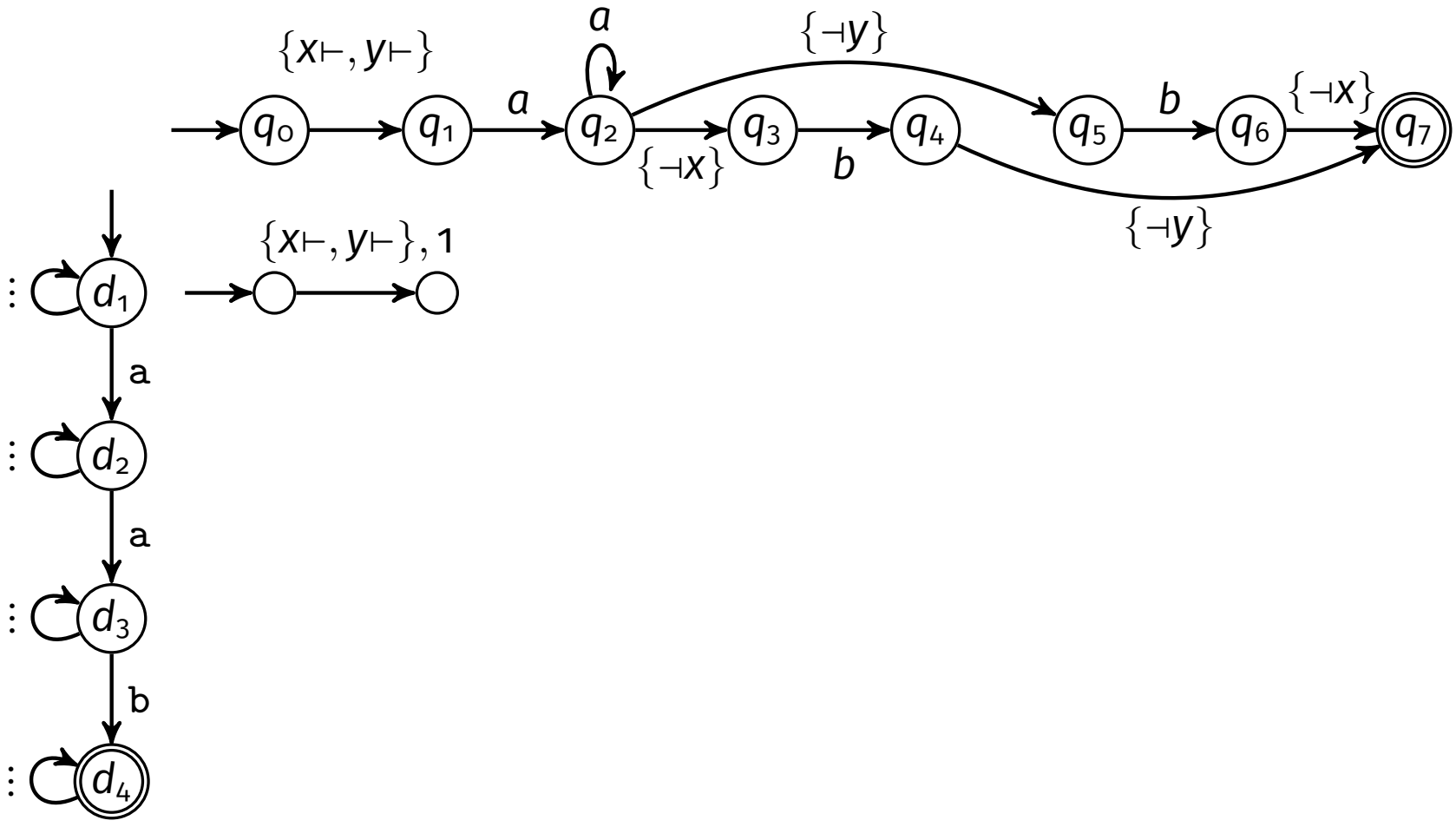

$\{x\vdash, y\vdash\}$
a
$\{\dashv y\}$
q_0
q_1
a
q_2
q_3
q_4
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv x\}$
b
$\{\dashv y\}$


d_1
a
d_2
a
d_3
b
d_4

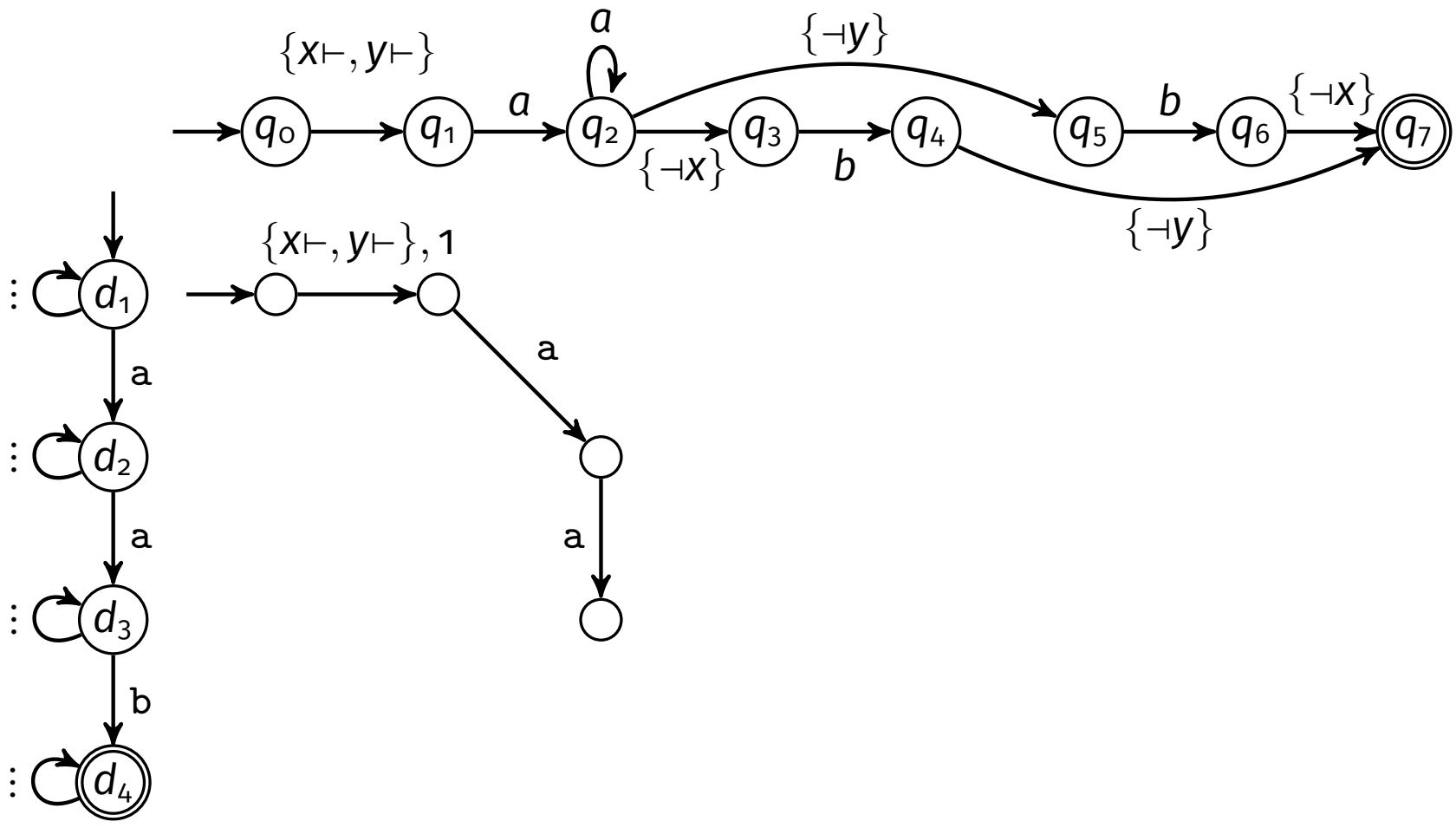
Idea


$\{x\vdash, y\vdash\}$
a
$\{\dashv y\}$
q_0
q_1
a
q_2
q_3
q_4
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv x\}$
b
$\{\dashv y\}$
d_1
a
d_2
a
d_3
b
d_4

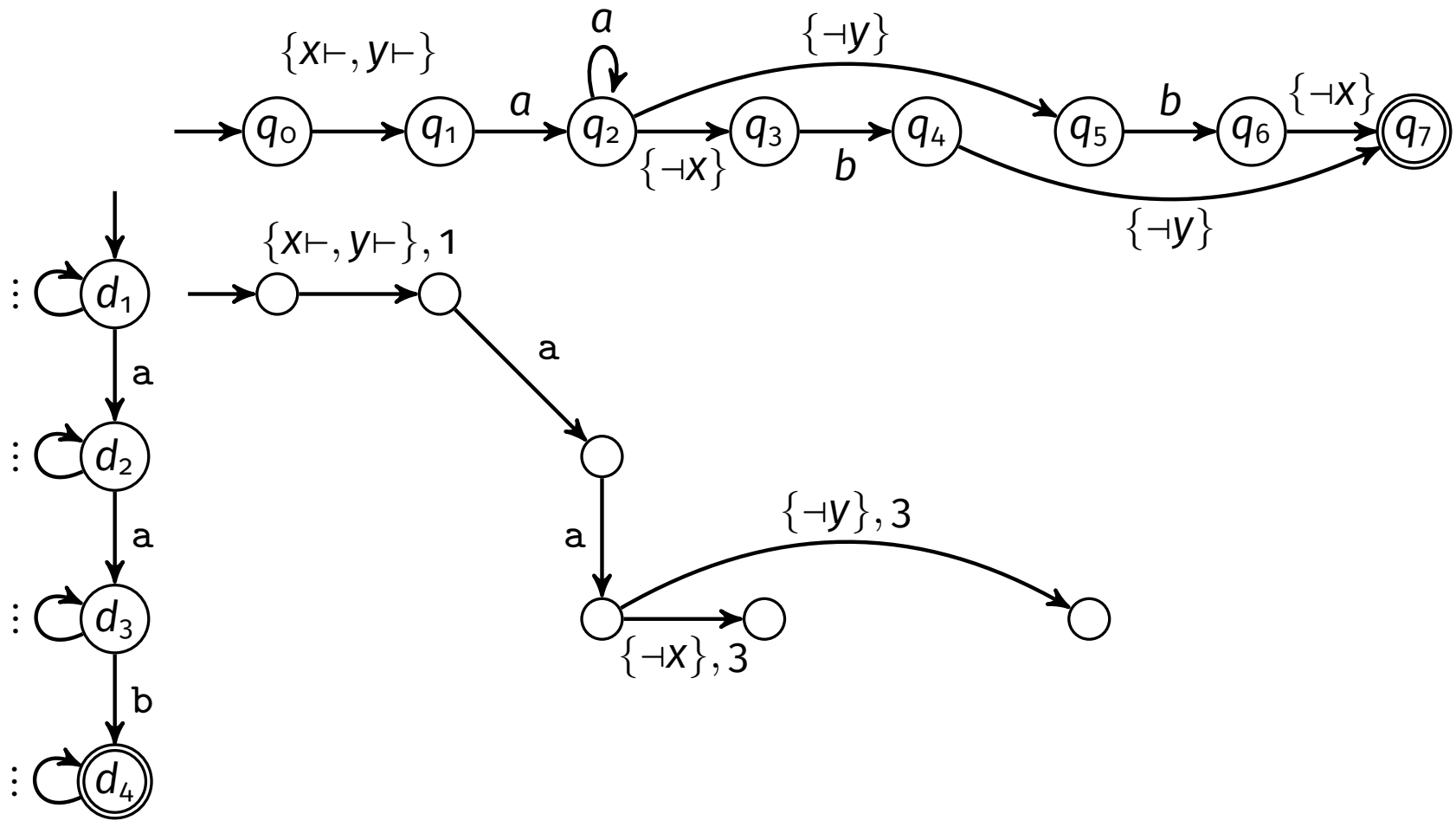
Idea


$\{x\vdash, y\vdash\}$
a
$\{\dashv y\}$
q_0
q_1
a
q_2
q_3
q_4
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv x\}$
b
$\{\dashv y\}$
$\{x\vdash, y\vdash\}, 1$
d_1
a
d_2
a
d_3
b
d_4

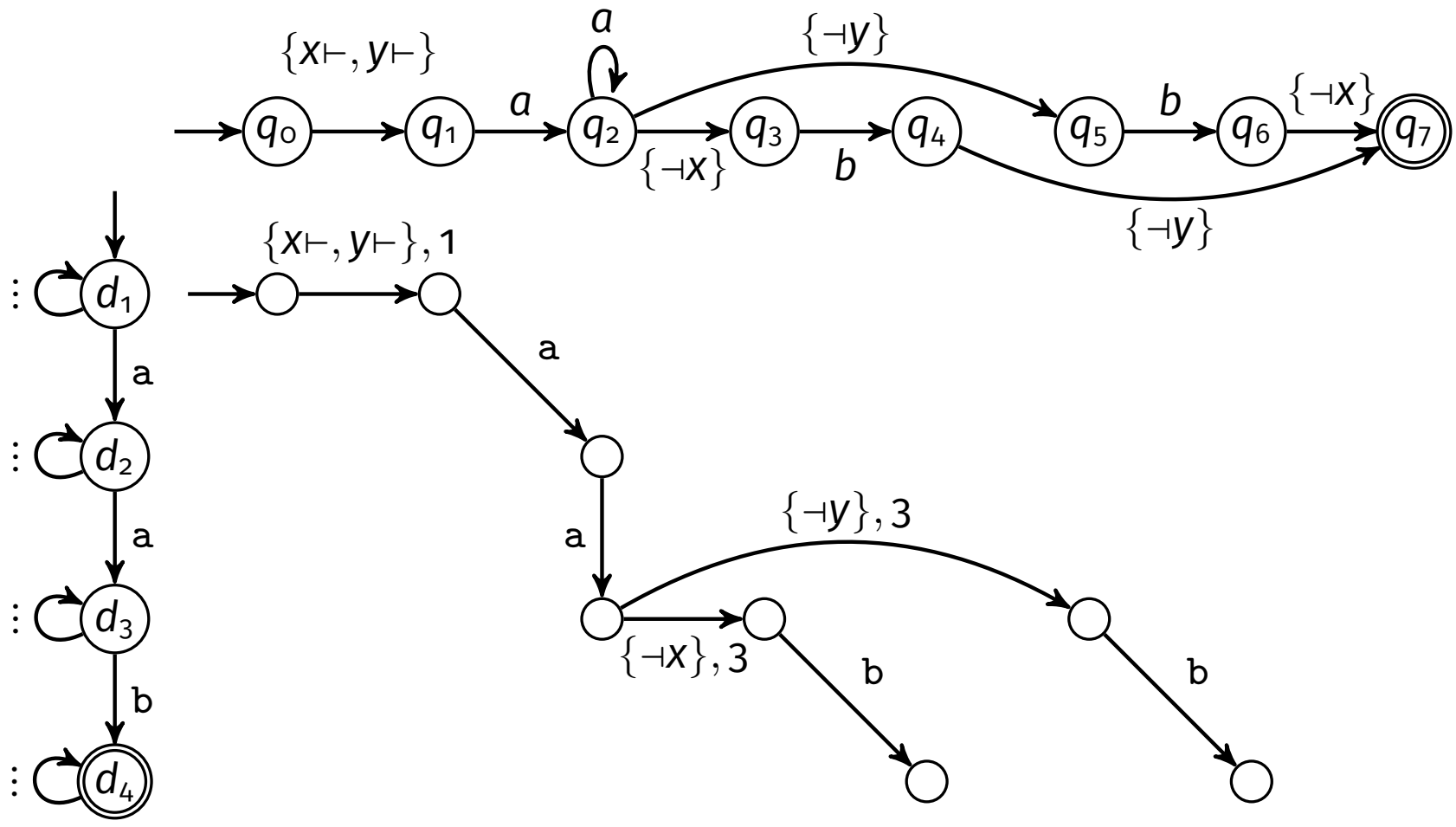
Idea


$\{x\vdash, y\vdash\}$
a
$\{\dashv y\}$
q_0
q_1
a
q_2
q_3
q_4
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv x\}$
b
$\{\dashv y\}$
$\{x\vdash, y\vdash\}, 1$
d_1
a
d_2
a
d_3
b
d_4
a
a

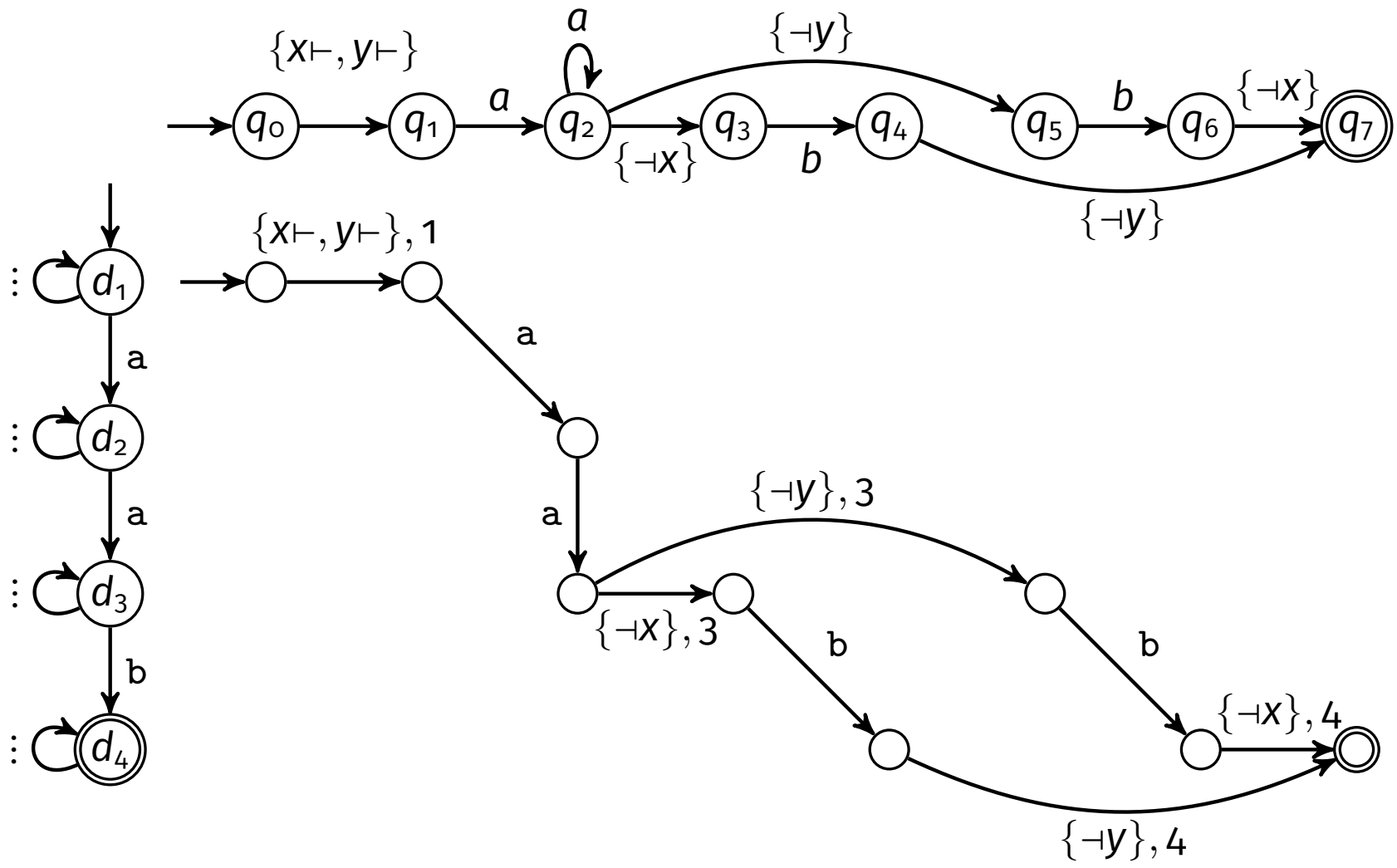
Idea


$\{x\vdash, y\vdash\}$
q_0
q_1
a
q_2
a
$\{\dashv x\}$
q_3
b
q_4
$\{\dashv y\}$
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv y\}$
$\{x\vdash, y\vdash\}, 1$
d_1
a
d_2
a
d_3
b
d_4
a
a
$\{\dashv y\}, 3$
$\{\dashv x\}, 3$

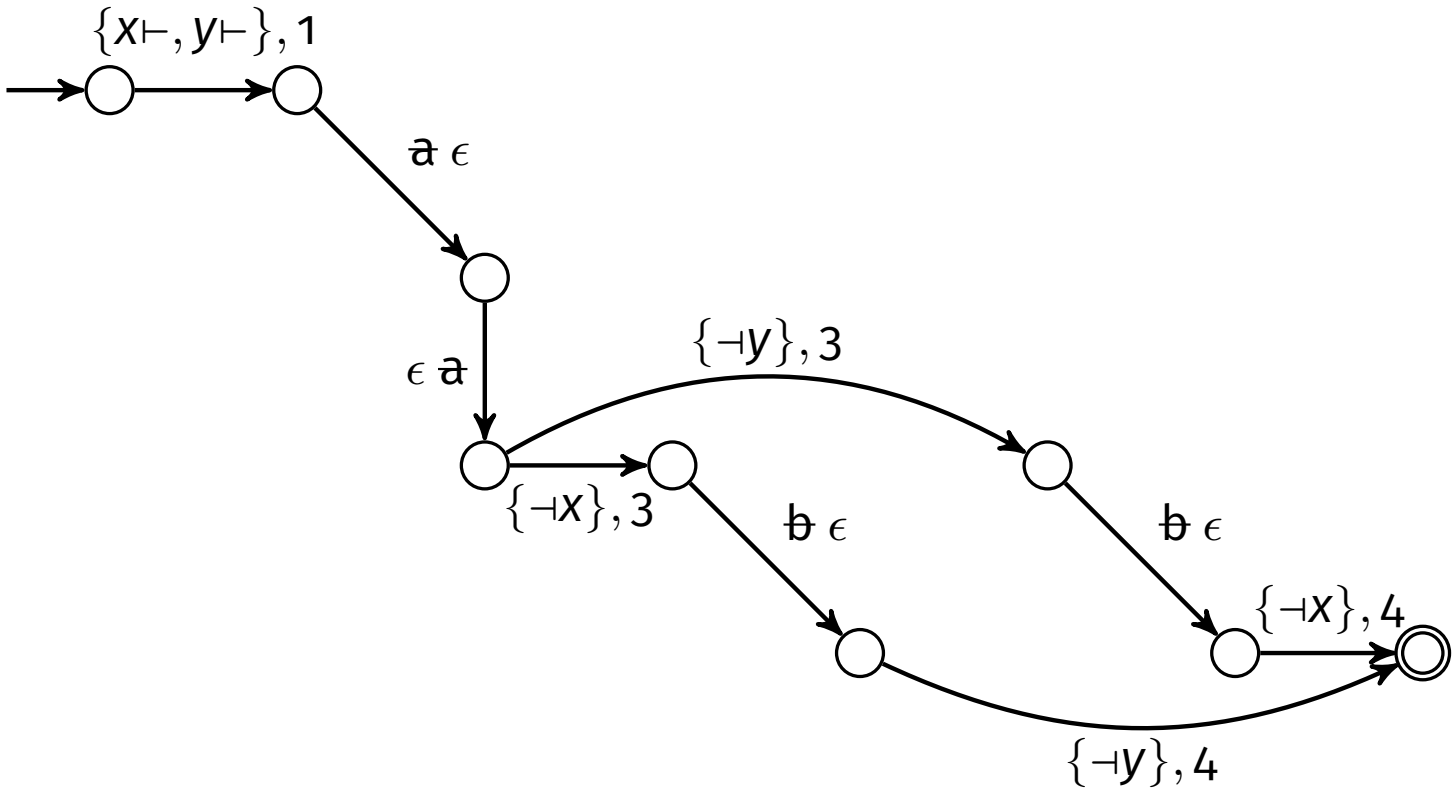
Idea


$\{x\vdash, y\vdash\}$
q_0
q_1
a
q_2
a
$\{\dashv x\}$
q_3
b
q_4
$\{\dashv y\}$
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv y\}$
$\{x\vdash, y\vdash\}, 1$
d_1
a
d_2
a
d_3
b
d_4
a
a
$\{\dashv y\}, 3$
$\{\dashv x\}, 3$
b
b

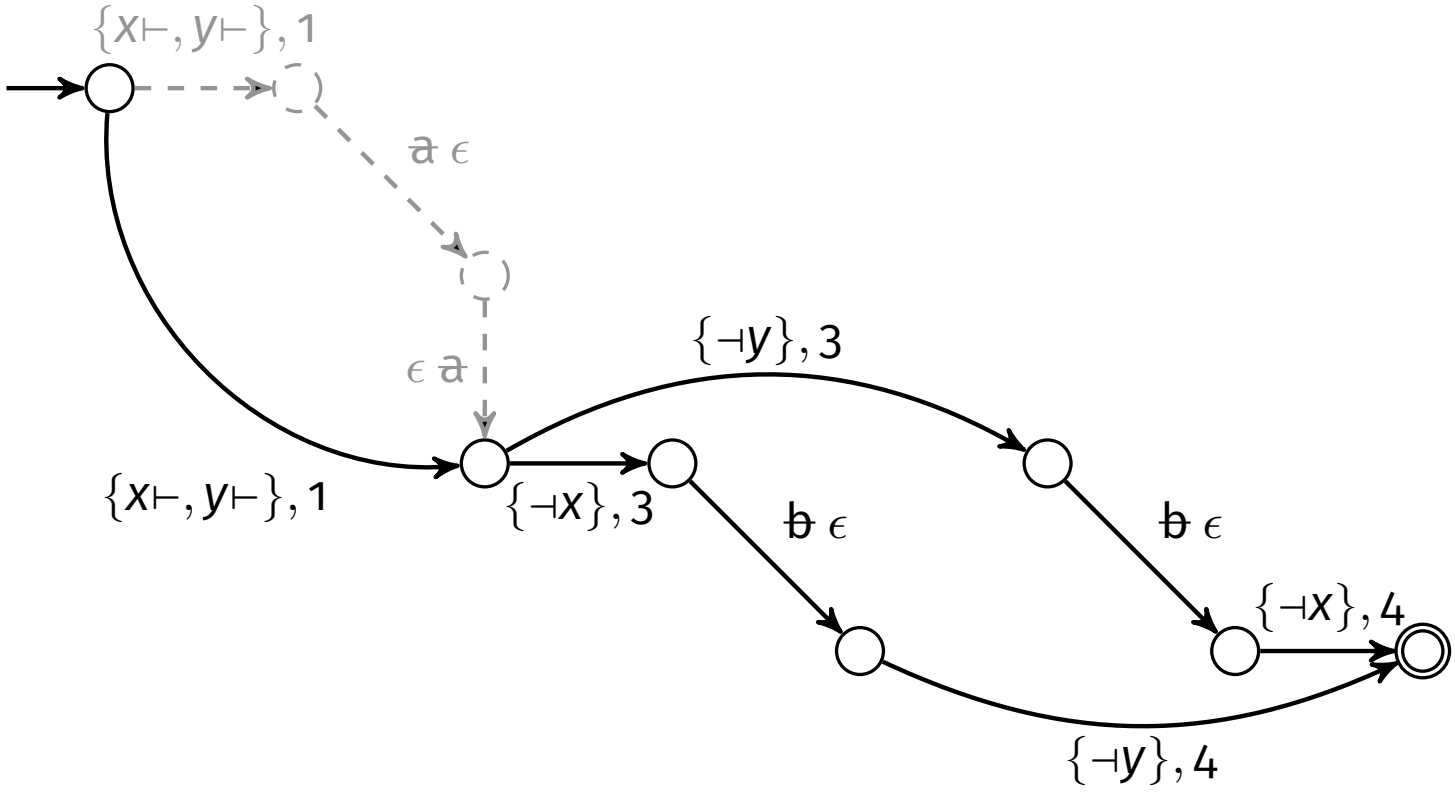
Idea


$\{x\vdash, y\vdash\}$
q_0
q_1
a
q_2
a
$\{\dashv x\}$
q_3
b
q_4
$\{\dashv y\}$
q_5
b
q_6
$\{\dashv x\}$
q_7
$\{\dashv y\}$
d_1
a
d_2
a
d_3
b
d_4
$\{x\vdash, y\vdash\}, 1$
a
a
$\{\dashv y\}, 3$
$\{\dashv x\}, 3$
b
b
$\{\dashv x\}, 4$
$\{\dashv y\}, 4$

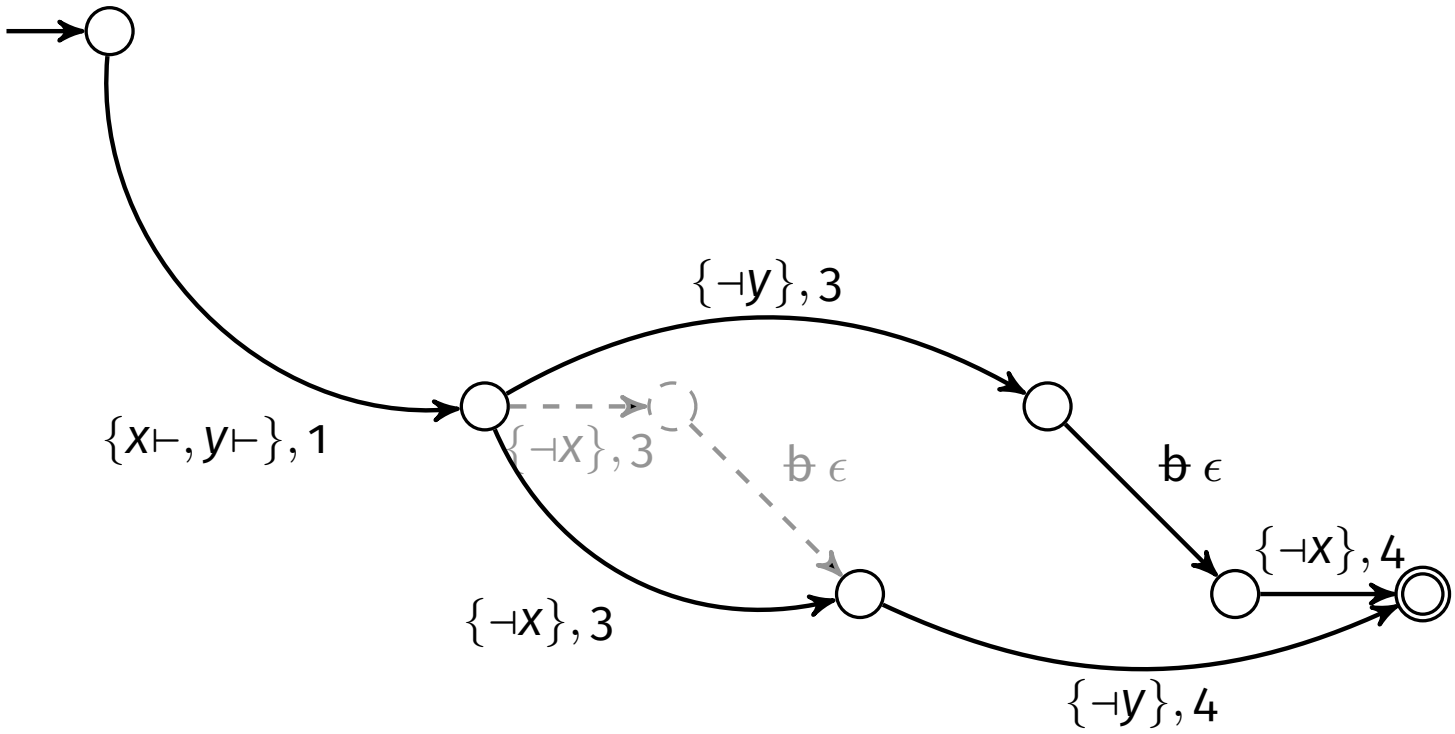
Idea


$\{x\vdash, y\vdash\}, 1$
~~a~~ ϵ
ϵ ~~a~~
$\{\neg y\}, 3$
$\{\neg x\}, 3$
~~b~~ ϵ
~~b~~ ϵ
$\{\neg x\}, 4$
$\{\neg y\}, 4$

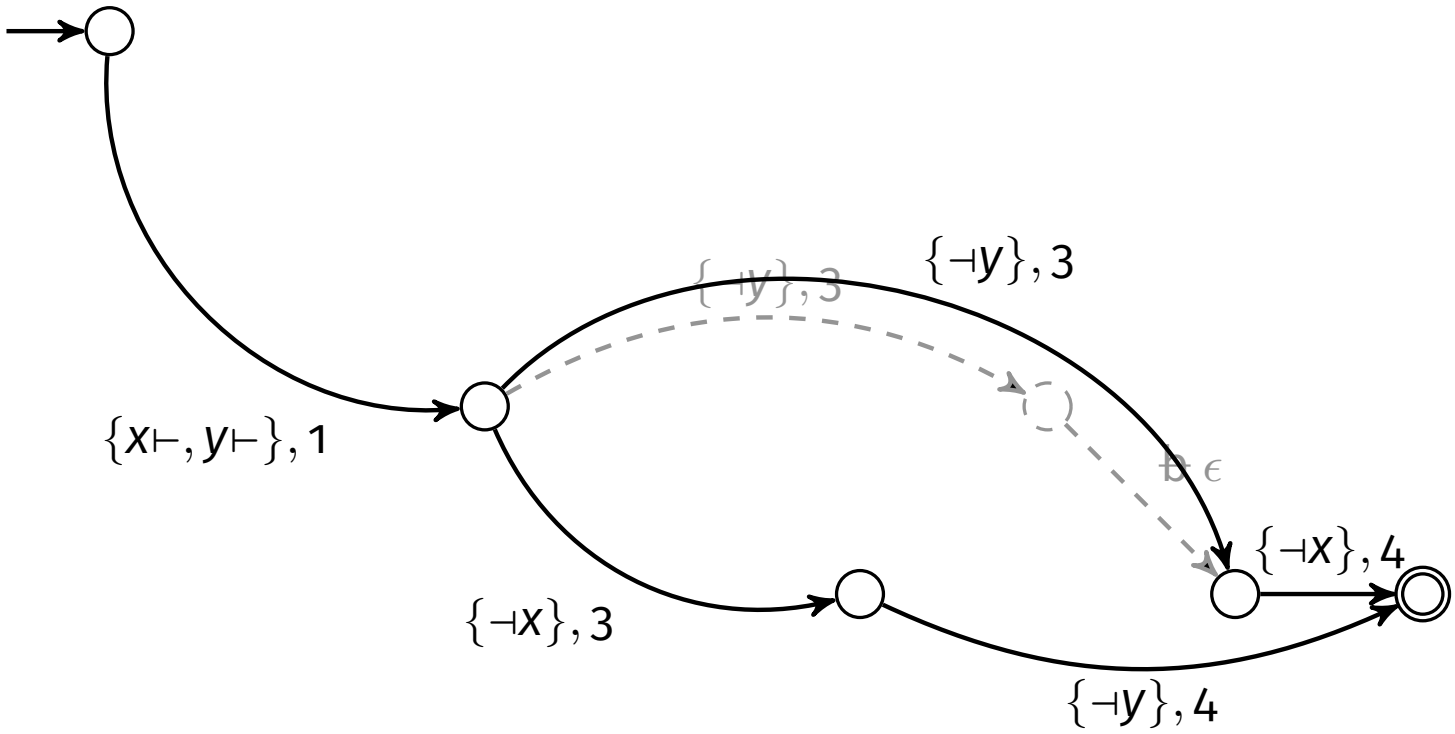
Idea


$\{x\vdash, y\vdash\}, 1$
$\mathtt{a}\ \epsilon$
$\epsilon\ \mathtt{a}$
$\{x\vdash, y\vdash\}, 1$
$\{\neg\vdash y\}, 3$
$\{\neg\vdash x\}, 3$
$\mathtt{b}\ \epsilon$
$\mathtt{b}\ \epsilon$
$\{\neg\vdash x\}, 4$
$\{\neg\vdash y\}, 4$

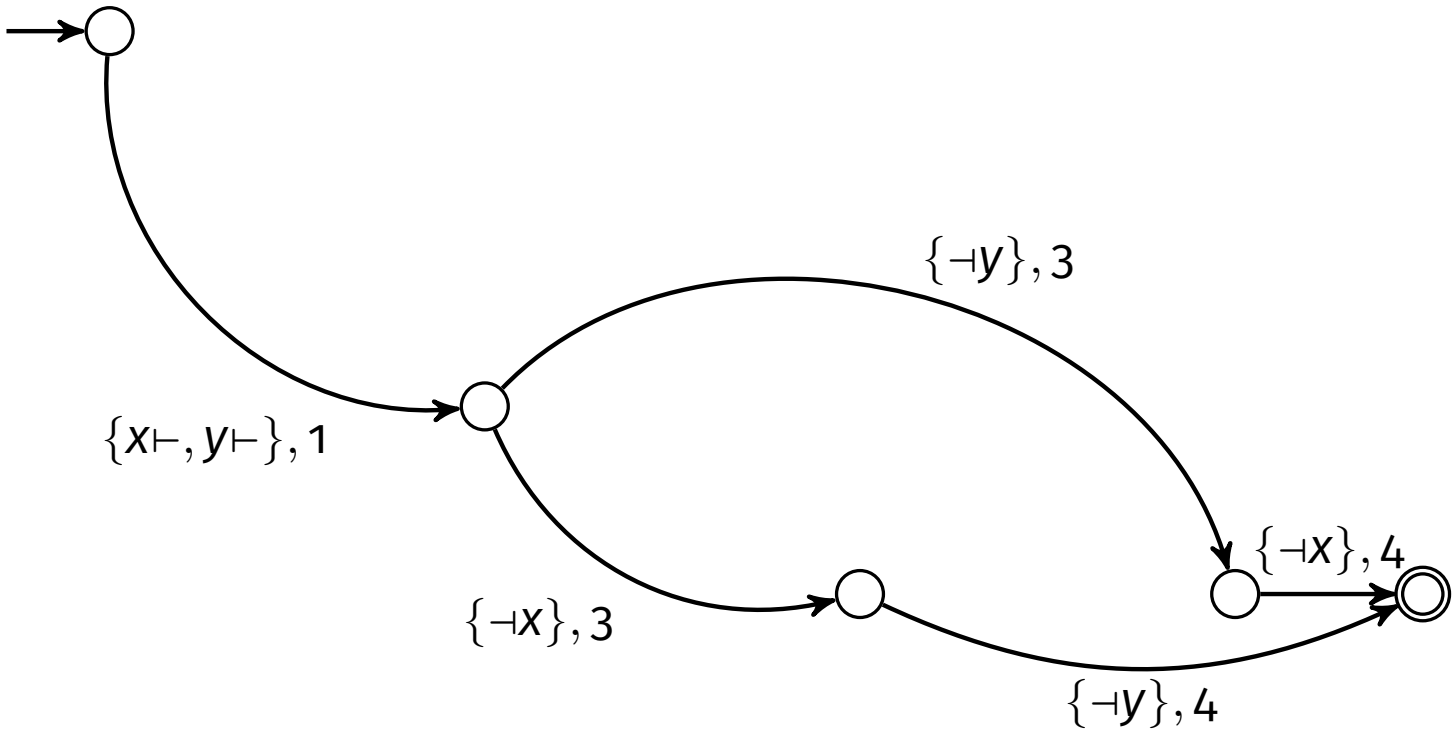
Idea


$\{\neg y\}, 3$
$\{x\vdash, y\vdash\}, 1$
$\{\neg x\}, 3$
$\not{b}\ \epsilon$
$\not{b}\ \epsilon$
$\{\neg x\}, 4$
$\{\neg x\}, 3$
$\{\neg y\}, 4$

Idea


$\{x\vdash, y\vdash\}, 1$
$\{\dashv y\}, 3$
$\{\dashv y\}, 3$
$\{\dashv x\}, 3$
ϵ
$\{\dashv x\}, 4$
$\{\dashv y\}, 4$

Idea



Desarrollo

Desarrollo

Mis tareas:

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.
- Buscar posibles errores en las demostraciones.

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.
- Buscar posibles errores en las demostraciones.
- Escribir en **LaTex**...

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.
- Buscar posibles errores en las demostraciones.
- Escribir en **LaTex**...
- Escribir en **LaTex**...

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.
- Buscar posibles errores en las demostraciones.
- Escribir en **LaTex**...
- Escribir en **LaTex**...
- Escribir en **LaTex**...

Desarrollo

Mis tareas:

- Reuniones semanales donde discutiamos avances del algoritmo.
- Apoyar en la demostración matemática de teoremas.
- Buscar posibles errores en las demostraciones.
- Escribir en **LaTex**...
- Escribir en **LaTex**...
- Escribir en **LaTex**...

Cosas que **sí** aprendí en los cursos **Matemáticas Discretas** y **Teoría de Autómatas**, pero las profundicé.

Resultados

Constant Delay Algorithms for Regular Document Spanners

Fernando Florenzano
PUC Chile
faflorenzano@uc.cl

Cristian Riveros
PUC Chile
cristian.riveros@uc.cl

Martín Ugarte
Université Libre de Bruxelles
mugartec@ulb.ac.be

Stijn Vansummeren
Université Libre de Bruxelles
stijn.vansummeren@ulb.ac.be

Domagoj Vrgoč
PUC Chile
dvrgoc@ing.puc.cl

ABSTRACT

Regular expressions and automata models with capture variables are core tools in rule-based information extraction. These formalisms, also called regular document spanners, use regular languages in order to locate the data that a user wants to extract from a text document, and then store this data into variables. Since document spanners can easily generate large outputs, it is important to have good evaluation algorithms that can generate the extracted data in a quick succession, and with relatively little precomputation time. Towards this goal, we present a practical evaluation algorithm that allows constant delay enumeration of a spanner's output after a precomputation phase that is linear in the document. While the algorithm assumes that the spanner is specified in a syntactic variant of variable set automata, we also study how it can be applied when the spanner is specified by general variable set automata, regex formulas, or spanner algebras. Finally, we study the related problem of counting the number of outputs of a document spanner, providing a fine grained analysis of the classes of document spanners that support efficient enumeration of their results.

1 INTRODUCTION

Information extraction (IE for short) has recently received a fair amount of attention from the the database community. The introduction of rule-based IE [7, 9, 15] has revealed interesting connections with logic [11, 12], automata [9, 17], datalog programs [3, 22], and relational languages [6, 13, 16]. In rule-based IE, documents from which we extract the information are modelled as strings. This is a natural assumption for many formats in use today (e.g. JSON and XML files, CSV documents, or plain text). The extracted data are represented by *spans*. These are intervals inside the document string that record the start and end position of the extracted data, plus the substring (the data) that this interval spans. The process of information extraction can then be abstracted by the notion of *document spanners* [9]: operators that map strings to tuples containing spans.

The most basic way of defining document spanners is to use some form or regular expressions or automata with capture variables. The idea is that a regular language is used in order to locate the data to be extracted, and variables to store this data. This approach to IE has been widely adopted in the database literature [3, 9–11, 17],

Resultados

Contenidos

Volvamos a nuestras frases

Volvamos a nuestras frases

“No he hecho **suficientes cursos** como para investigar”

“No sé **en que** investigaría”

“No tengo **tiempo** para investigar”

“¿En que aportaría? **Está todo resuelto**”

“¿Qué gano yo? **Son solo créditos**”

“¿Y si **no resulta**?”

“Es terrible **fome** investigar”

“No he hecho suficientes cursos como para investigar”

“No he hecho suficientes cursos como para investigar”

Lo que necesitas **saber** por investigación **varia mucho**.
¡A veces solo necesitas un concepto del cual partir!

“No sé en que investigaría”

“No sé en que investigaría”

No tienes porque saber.

“No tengo tiempo para investigar”

“No tengo tiempo para investigar”

Nadie lo tiene.

“No tengo tiempo para investigar”

Nadie lo tiene. Siguen siendo alumnos y los cursos son primero, pero uno aprende a hacerse el tiempo.

“¿En que aportaría? Está todo resuelto”

“¿En que aportaría? Está todo resuelto”

En los cursos universitarios vemos el **límite del conocimiento**, hay preguntas sin responder hasta en los **cursos básicos**.

“¿Y si no resulta?”

“¿Y si no resulta?”

¡Se **intenta** de nuevo, o se cambia de **dirección**!
Incluso “fallando” se **aprende** algo en el camino.

“Es terrible fome investigar”

“Es terrible fome investigar”

Tiene partes **fomes**, pero tiene muchas **entretenidas**.

“¿Qué gano yo? Son solo créditos”

“¿Qué gano yo? Son solo créditos”

Se ganan **herramientas técnicas** que te **diferencian**.

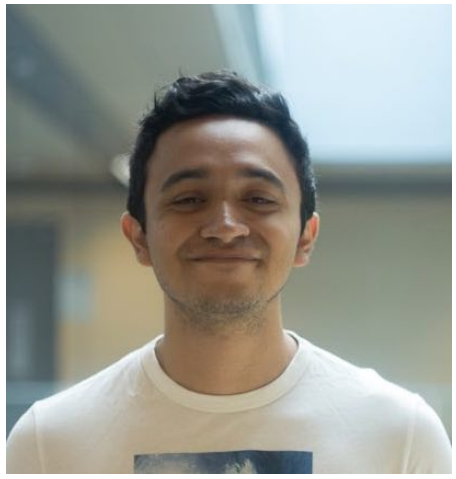
“¿Qué gano yo? Son solo créditos”

Se ganan **herramientas técnicas** que te **diferencian**.
Se aprende a trabajar en **equipo**.

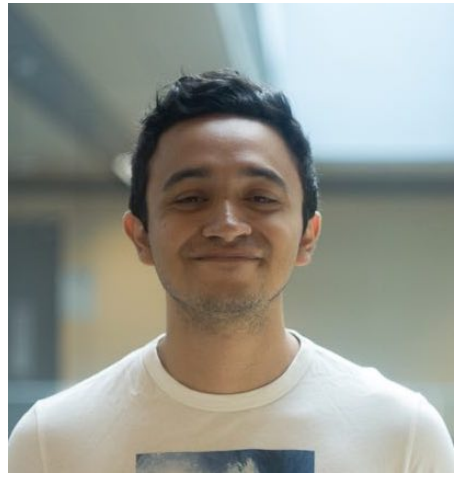
“¿Qué gano yo? Son solo créditos”

Se ganan **herramientas técnicas** que te **diferencian**.
Se aprende a trabajar en **equipo**.
Se aprende a ser **persona**.



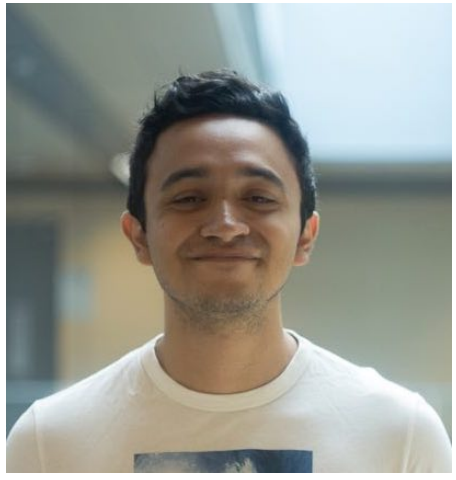


Si el **joven** Fernando pudo, ¡tú también!



Si el **joven** Fernando pudo, ¡tú también!

¡Muchas gracias!



Si el **joven** Fernando pudo, ¡tú también!

¡Muchas gracias!

¿Alguna pregunta?

Agradecimientos

- Por código base de presentación, a **Cristian Riveros**.
- Por los aprendizajes **Juan Reutter**, **Cristian Riveros** y **Domagoj Vrgoč**.