Dependable Data Repairing with Fixing Rules

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

Jiannan Wang
Nan Tang

Data is Dirty

incomplete
inconsistent
inaccurate
...

Data is Dirty

incomplete
inconsistent
inaccurate
...

25% companies: flawed data
3+ trillion $: US economy
20%: labor productivity
... ...

Data is Dirty

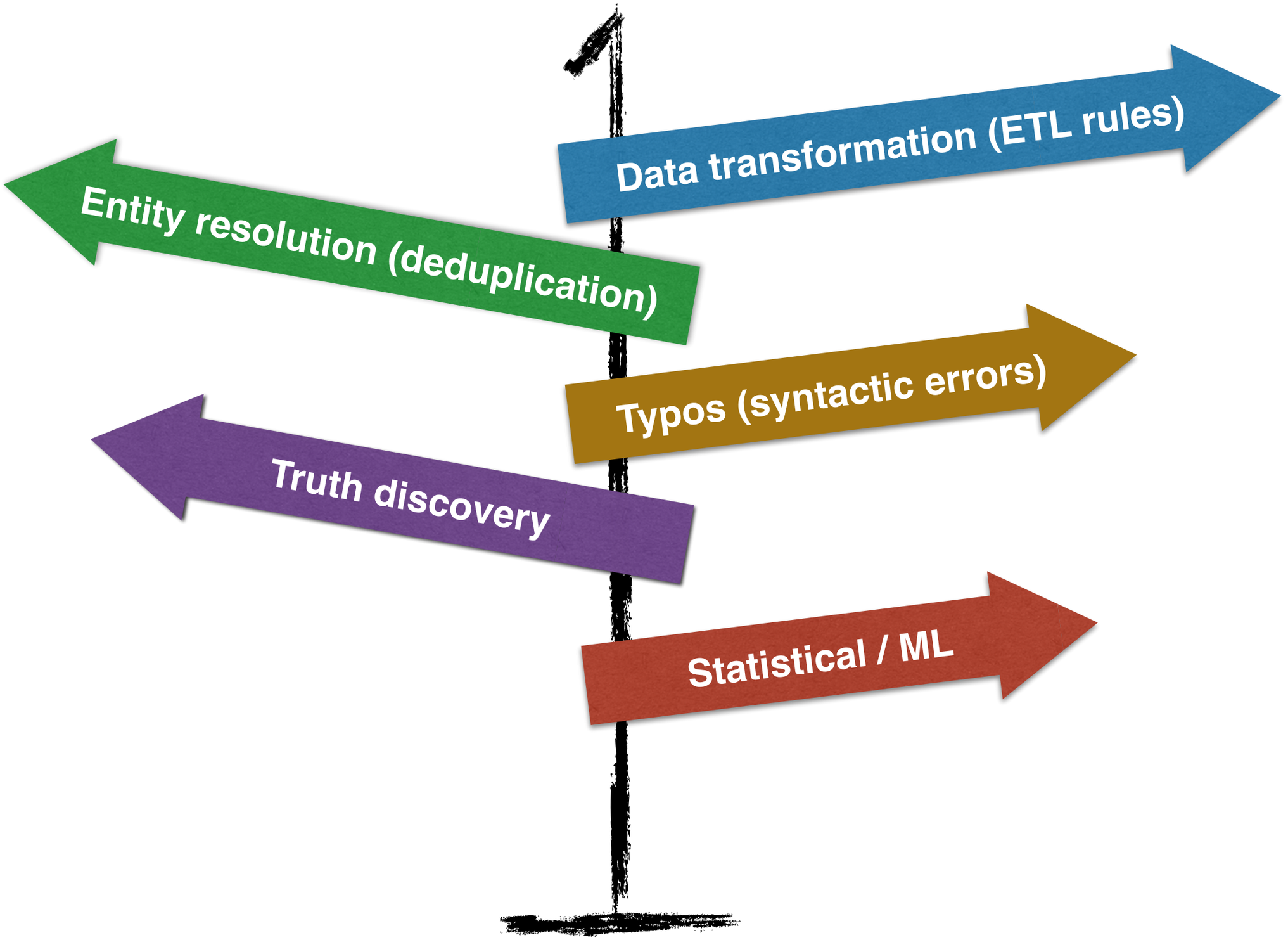
incomplete
inconsistent
inaccurate
...

25% companies: flawed data
3+ trillion $: US economy
20%: labor productivity
... ...

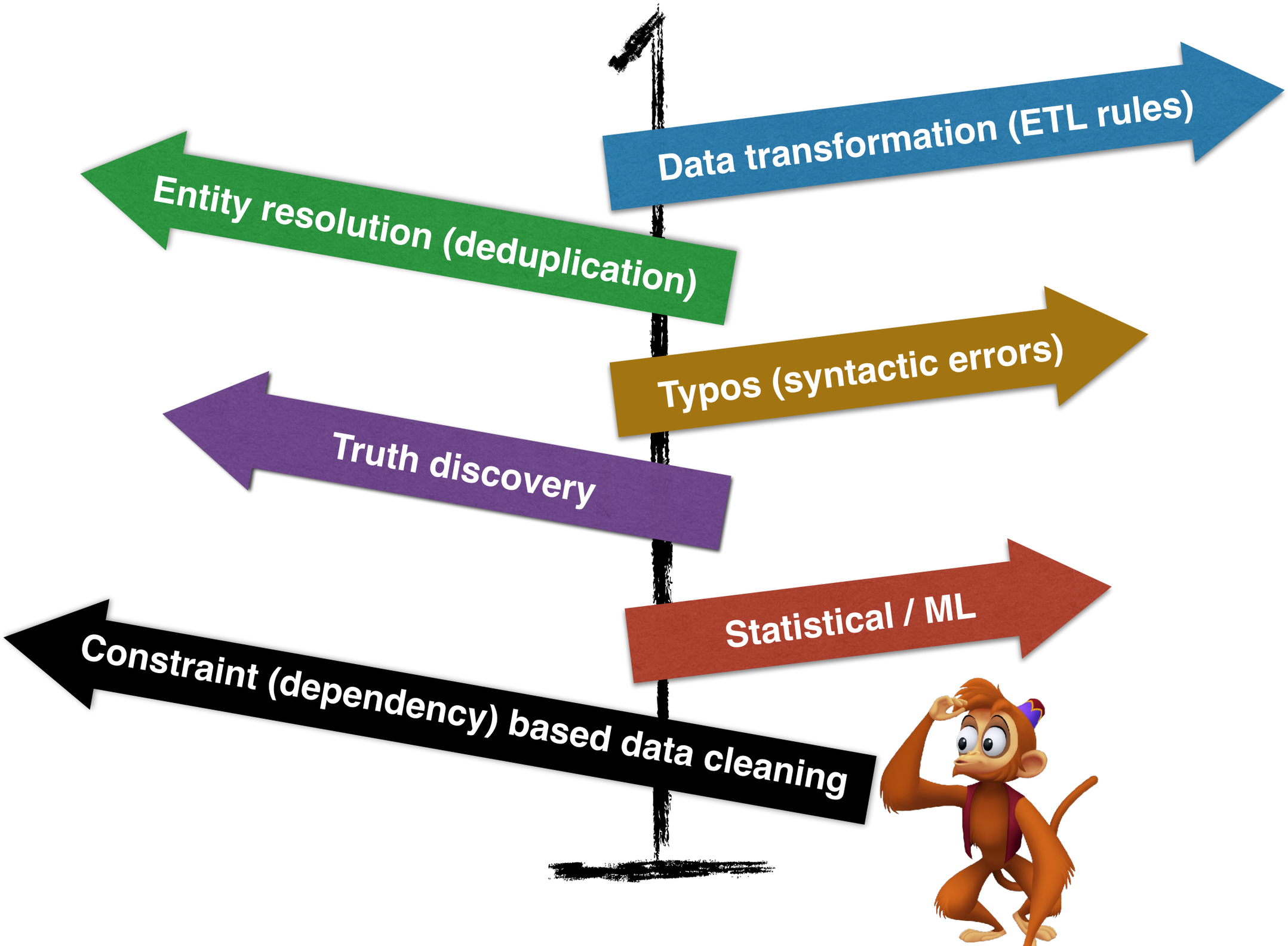
Data is Dirty

Big (clean) data: new oil

State-of-the-art



State-of-the-art



Dependency Theory

- Data dependencies (*a.k.a.* integrity constraints)

Dependency Theory

- Data dependencies (*a.k.a.* integrity constraints)

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Dependency Theory

- Data dependencies (*a.k.a.* integrity constraints)

FD: [country] -> [capital]

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Dependency Theory

- Data dependencies (*a.k.a.* integrity constraints)

FD: [country] -> [capital]

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Dependency Theory

- Data dependencies (*a.k.a.* integrity constraints)

FD: [country] -> [capital]

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Data dependencies are not sufficient to guide dependable data repairing

User Guidance

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

Is r2[country] China?
YES.

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

Is r2[country] China?
YES.

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

Is r2[country] China?
YES.

Is r1[country] China?

Is r3[country] China?

Is r4[country] Canada?

… … … …

User Guidance

editing rule: ((country, country) -> (capital, capital))

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

country	capital
China	Beijing
Canada	Ottawa
Japan	Tokyo

Is r2[country] China?
YES.

Is r1[country] China?

Is r3[country] China?

Is r4[country] Canada?

… … … …

check ***each*** tuple: not cheap !!

precision: +
recall: ++

Heuristic
(Automated)

precision: ++
recall: ++

Certain
(User guided)

precision: +
recall: ++

Heuristic (Automated)

precision: ++
recall: +

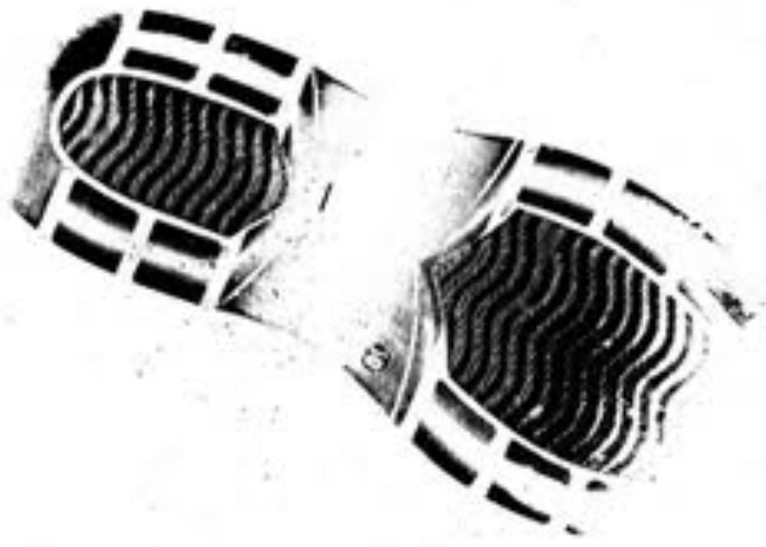
Fixing Rules (Automated)

precision: ++
recall: ++

Certain (User guided)

negative

evidence



negative

evidence

Data patterns

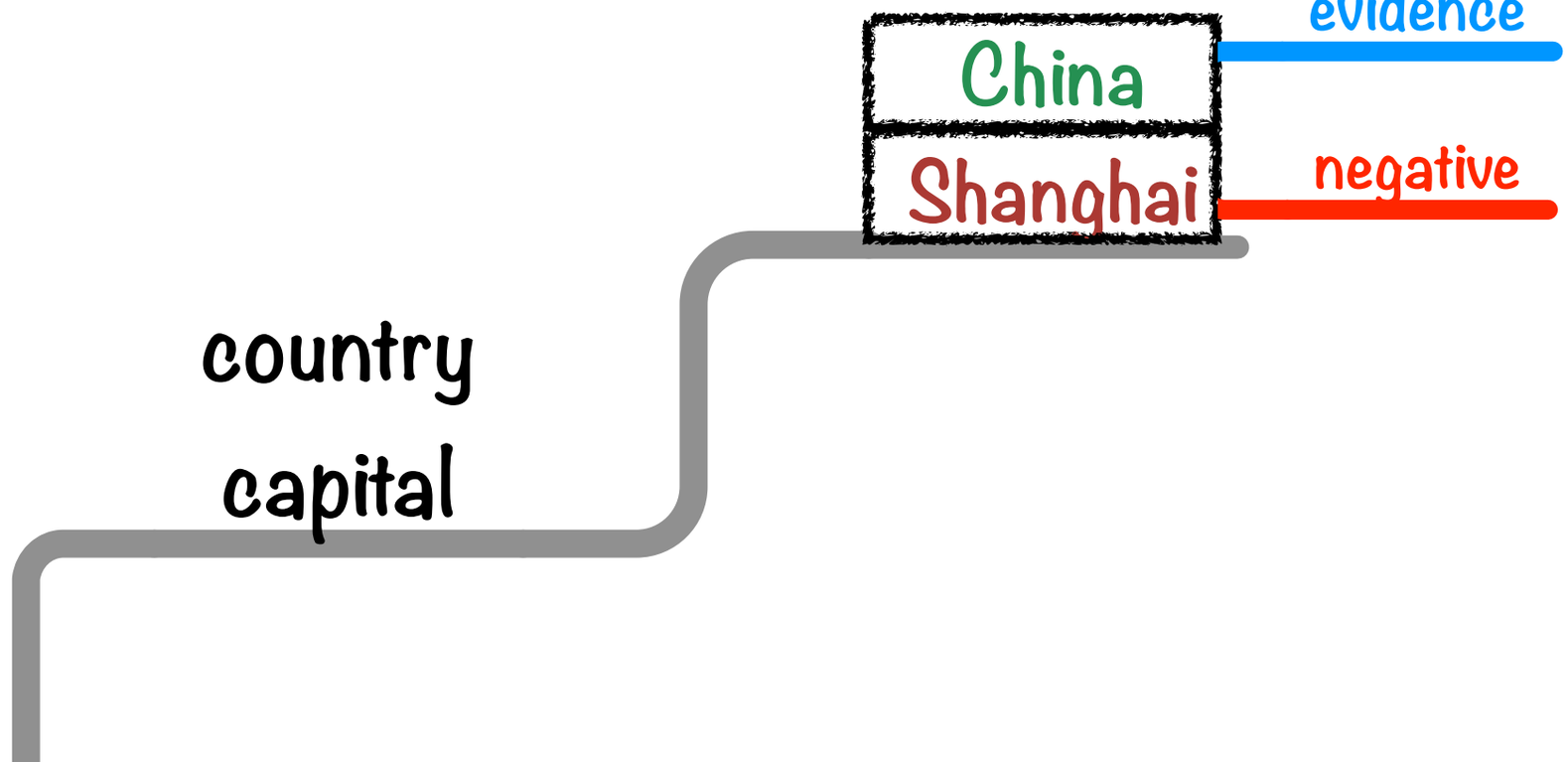
country

capital

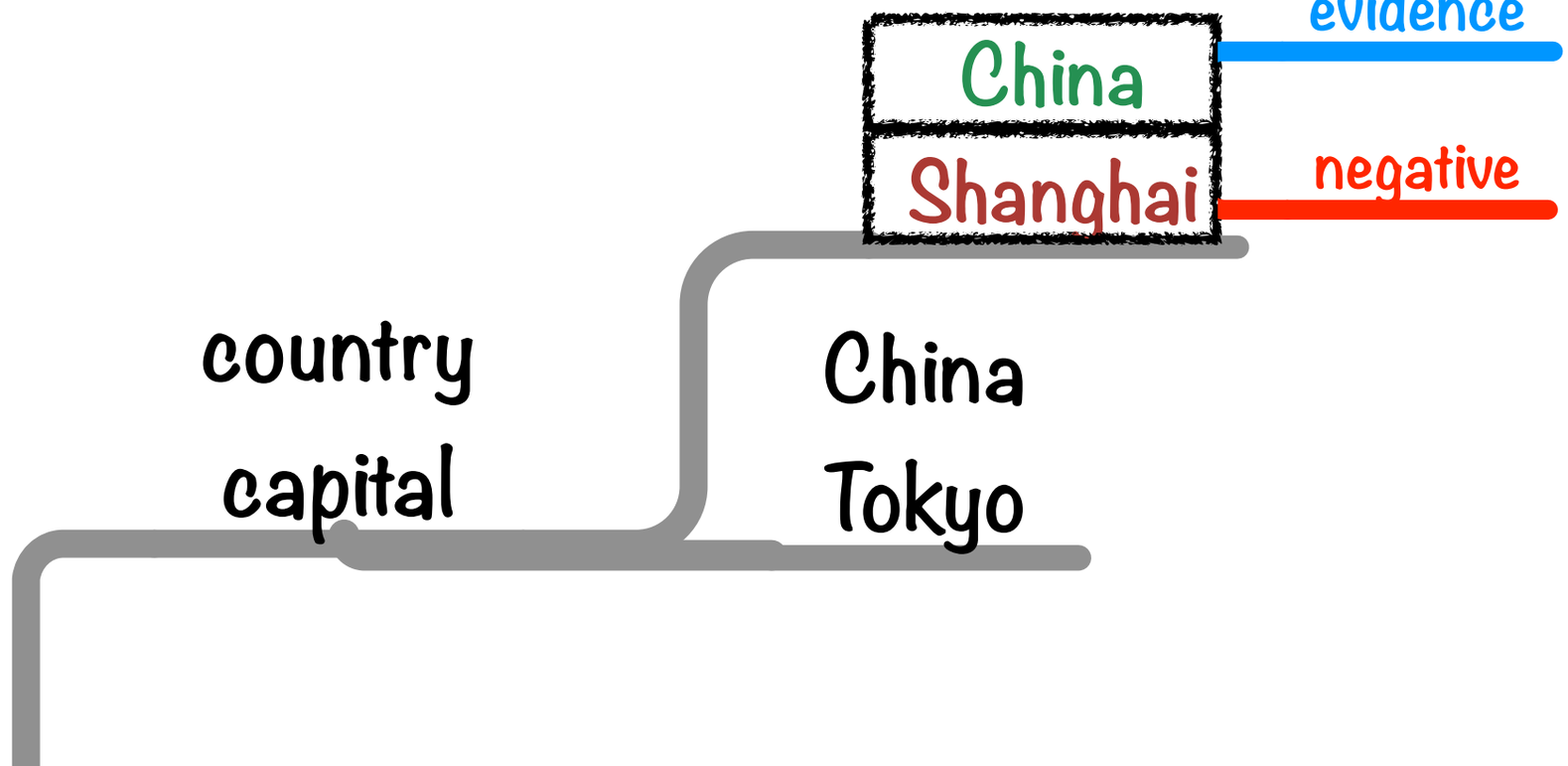
China

Shanghai

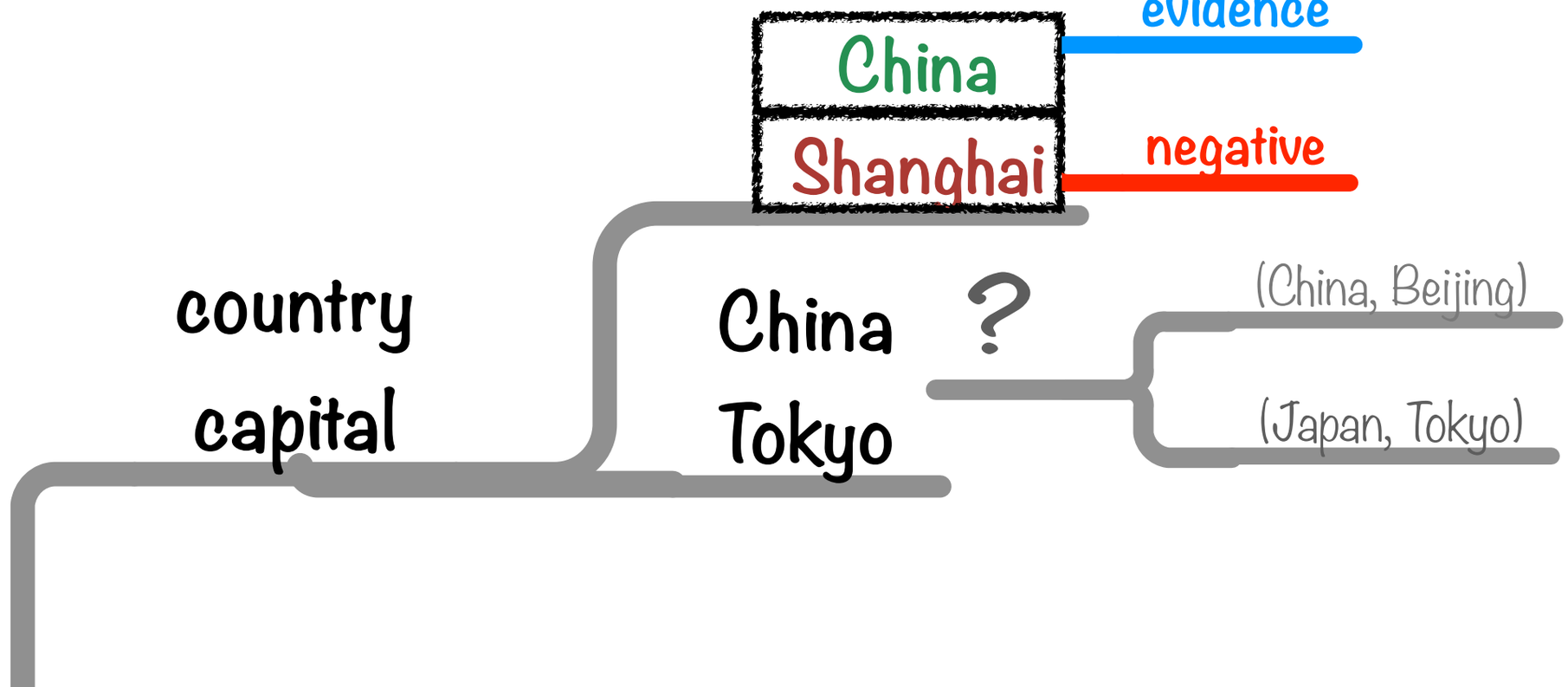
Data patterns

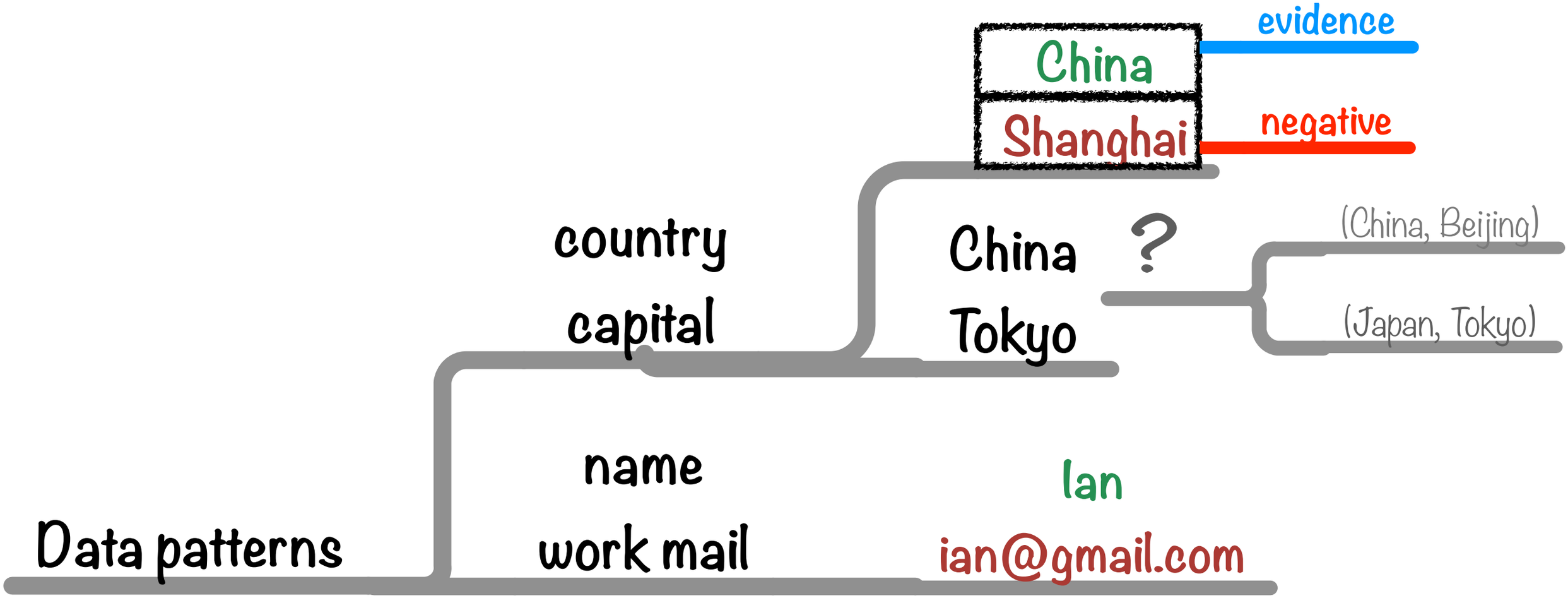


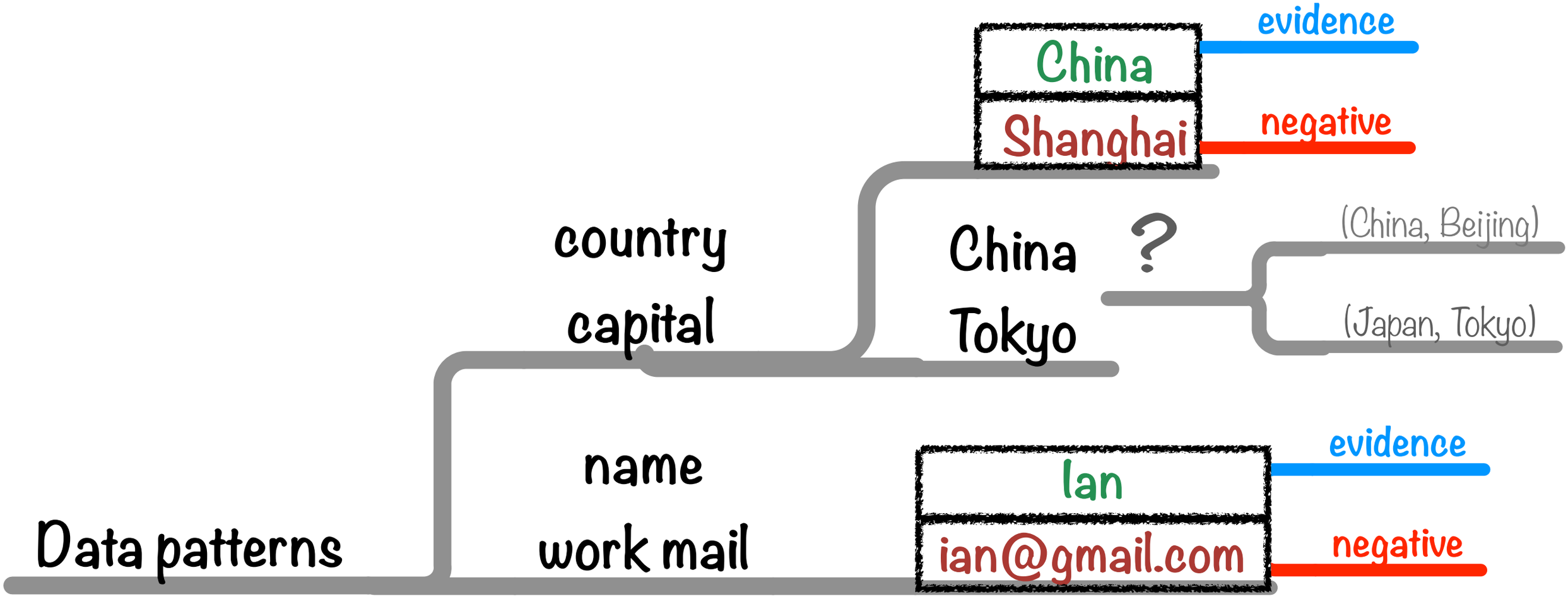
Data patterns

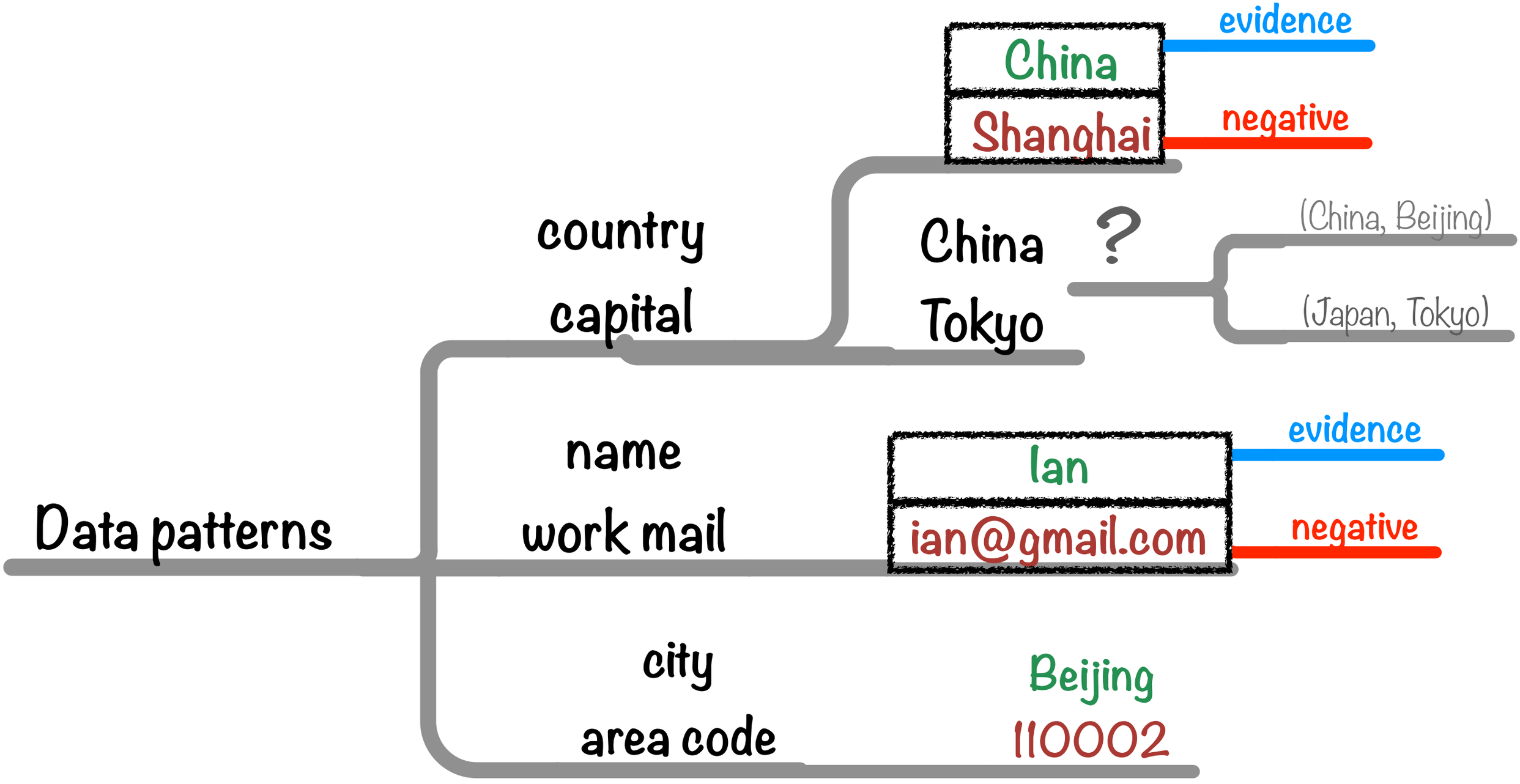


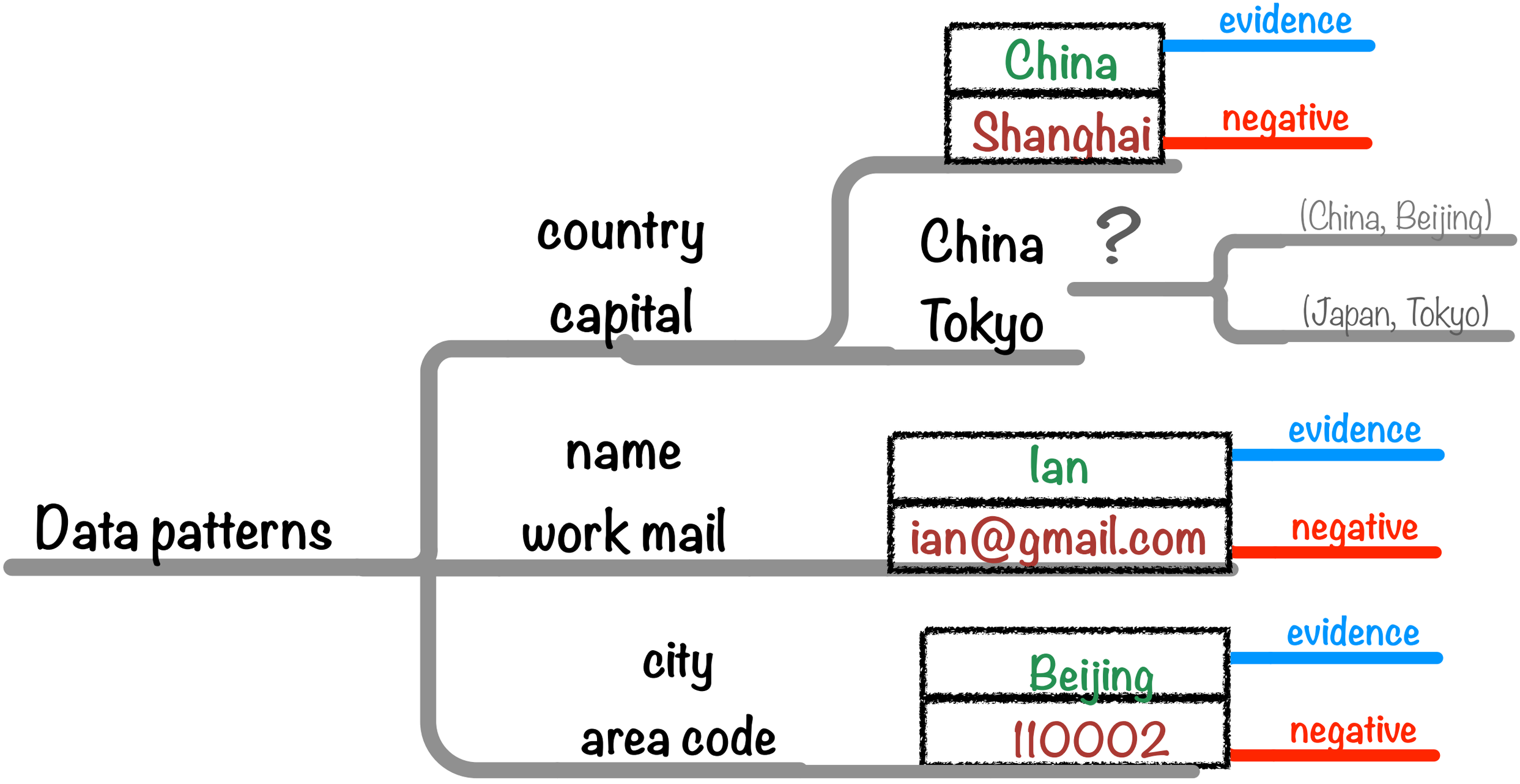
Data patterns




evidence
China
negative
Shanghai
country
capital
China
Tokyo
?
(China, Beijing)
(Japan, Tokyo)
name
work mail
lan
ian@gmail.com
Data patterns


country
capital
China
Shanghai
evidence
negative
China
Tokyo
?
(China, Beijing)
(Japan, Tokyo)
Data patterns
name
work mail
lan
ian@gmail.com
evidence
negative


country
capital
China
Shanghai
evidence
negative
China
Tokyo
?
(China, Beijing)
(Japan, Tokyo)
Data patterns
name
work mail
Ian
ian@gmail.com
evidence
negative
city
area code
Beijing
110002


Data patterns
country
capital
China
Tokyo
China
Shanghai
evidence
negative
China
?
(China, Beijing)
(Japan, Tokyo)
name
work mail
Ian
ian@gmail.com
evidence
negative
city
area code
Beijing
110002
evidence
negative

Fixing Rules

- **Syntax**

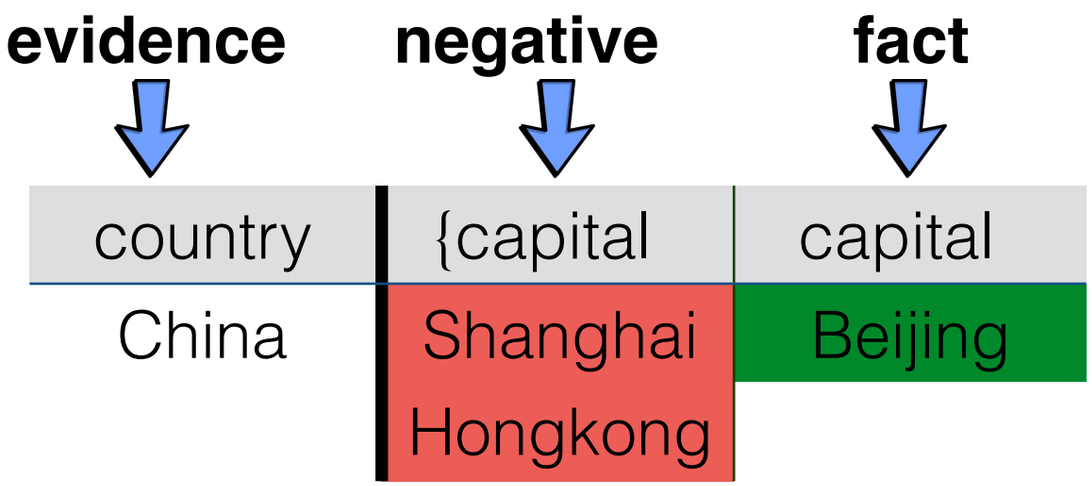
fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing

evidence	negative	fact
country	{capital	capital
China	Shanghai Hongkong	Beijing

Fixing Rules

- **Syntax**

fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing

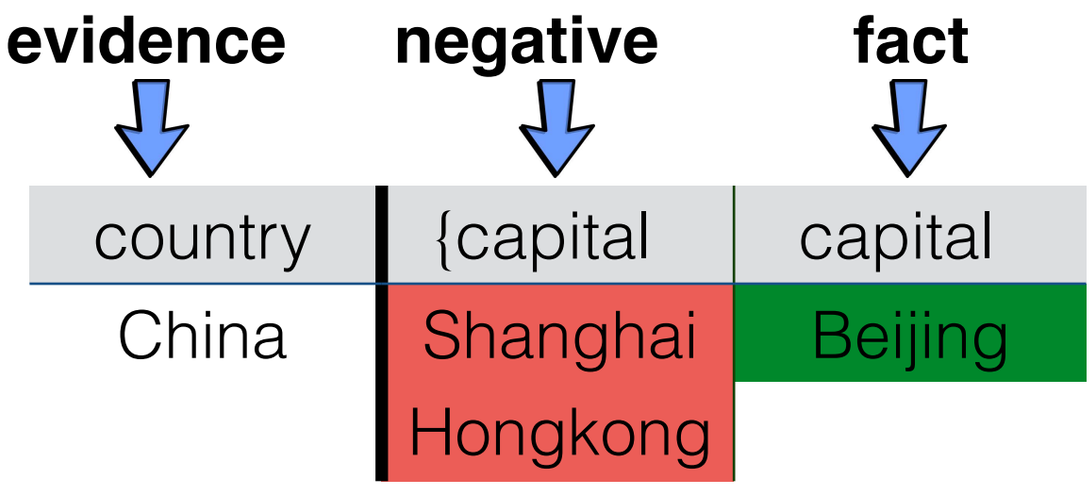


	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Fixing Rules

- **Syntax**

fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing



	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Fixing Rules

- **Syntax**

fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing

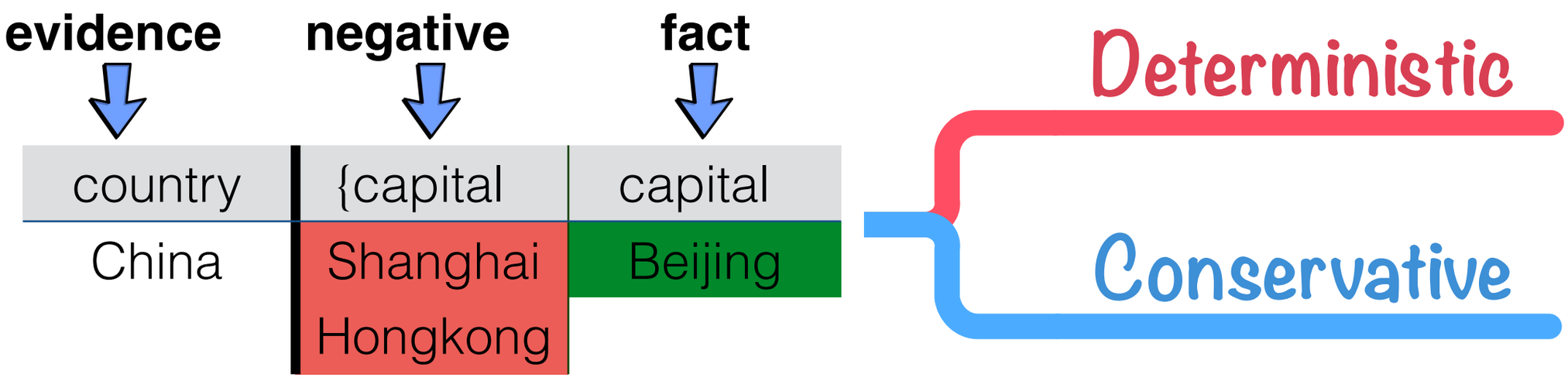
evidence	negative	fact
country	{capital	capital
China	Shanghai Hongkong	Beijing

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Beijing	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Fixing Rules

- **Syntax**

fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing



	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Beijing	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

Applying One Fixing Rule

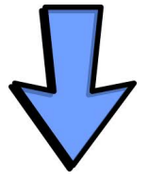
country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

r2	Ian	China	Shanghai	Hongkong	ICDE

Applying One Fixing Rule

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

r2	Ian	China	Shanghai	Hongkong	ICDE



r2’	Ian	China	Beijing	Hongkong	ICDE

Applying Multiple Fixing Rules

- **Fixes**

fR_1'

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	
	Tokyo	

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

Applying Multiple Fixing Rules

- **Fixes**

fR_1'

country	{capital	capital
China	Shanghai Hongkong Tokyo	Beijing

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

r2

Ian	China	Shanghai	Hongkong	ICDE

fR_1'

r2'

Ian	China	Beijing	Hongkong	ICDE

Applying Multiple Fixing Rules

- **Fixes**

fR_1'

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	
	Tokyo	

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

r2

Ian	China	Shanghai	Hongkong	ICDE

fR_1'

r2'

Ian	China	Beijing	Hongkong	ICDE

- **Unique fixes**

r3

Peter	China	Tokyo	Tokyo	ICDE

fR_1'

r3'

Peter	China	Beijing	Tokyo	ICDE

Applying Multiple Fixing Rules

- **Fixes**

fR_1'

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	
	Tokyo	

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

r2	Ian	China	Shanghai	Hongkong	ICDE

fR_1'

r2'	Ian	China	Beijing	Hongkong	ICDE

- **Unique fixes**

r3	Peter	China	Tokyo	Tokyo	ICDE

fR_1'

r3'	Peter	China	Beijing	Tokyo	ICDE

≠

fR_3

r3''	Peter	Japan	Tokyo	Tokyo	ICDE

Fundamentals

Fundamental Problems

Termination	Yes
Consistency	PTIME
Implication	coNP-complete
Determinism	Yes

Ensuring Consistency

- **Tuple enumeration**

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

Ensuring Consistency

- **Tuple enumeration**

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

(∘, China, Shanghai, ∘, ∘)
(∘, China, Hongkong, ∘, ∘)
(∘, China, Toronto, ∘, ∘)

(∘, Canada, Shanghai, ∘, ∘)
(∘, Canada, Hongkong, ∘, ∘)
(∘, Canada, Toronto, ∘, ∘)

Ensuring Consistency

- **Tuple enumeration**

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

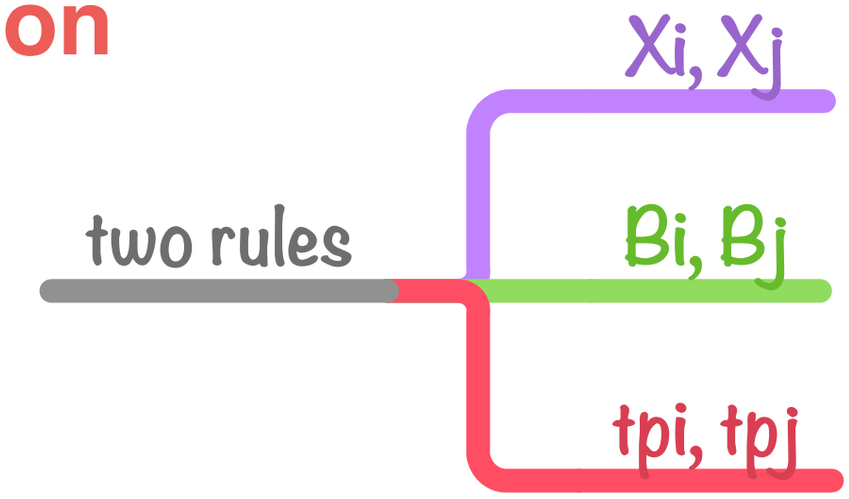
fR_2

country	{capital	capital
Canada	Toronto	Ottawa

(∘, China, Shanghai, ∘, ∘)
(∘, China, Hongkong, ∘, ∘)
(∘, China, Toronto, ∘, ∘)

(∘, Canada, Shanghai, ∘, ∘)
(∘, Canada, Hongkong, ∘, ∘)
(∘, Canada, Toronto, ∘, ∘)

- **Rule characterization**



Repair

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai Hongkong	Beijing

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Key	List
country, China	fR1
country, Canada	fR2
conf, ICDE	fR3, fR4
capital, Tokyo	fR3
city, Tokyo	fR3
capital, Beijing	fR4

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai Hongkong	Beijing

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Key	List
country, China	fR1
country, Canada	fR2
conf, ICDE	fR3, fR4
capital, Tokyo	fR3
city, Tokyo	fR3
capital, Beijing	fR4

r1: cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1}
itr1: r1’ = r1, rules = {}

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

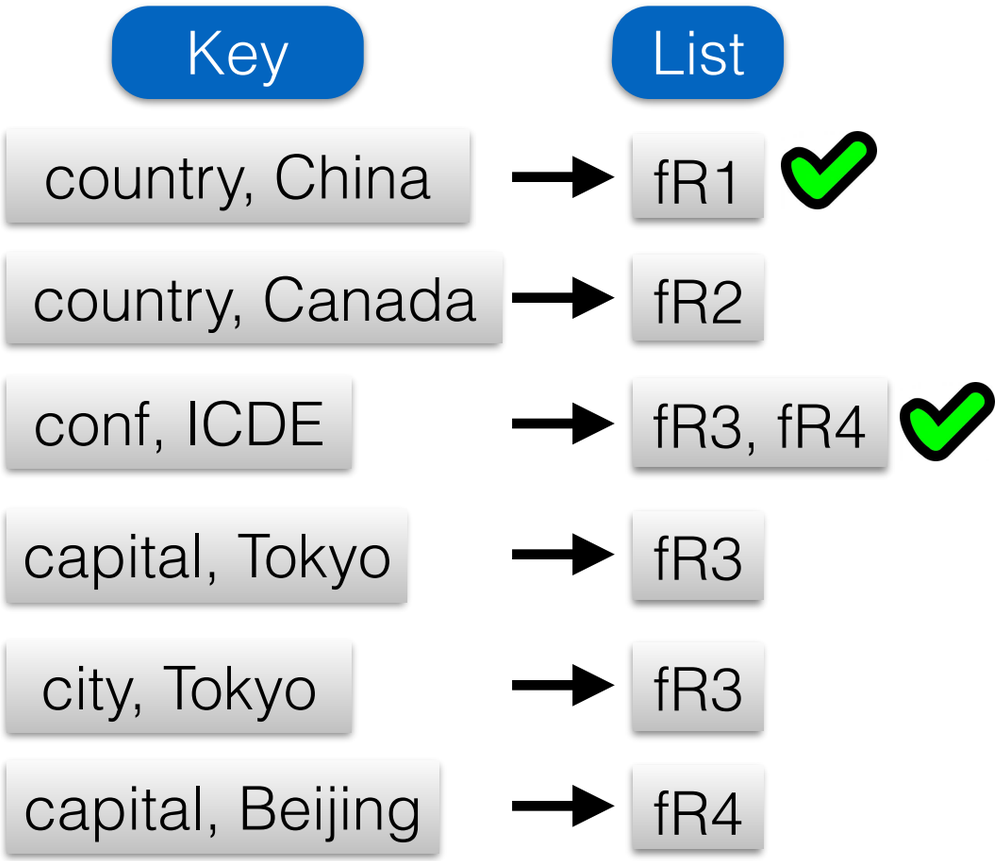
country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai



r1:
itr1:
cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1}
r1’ = r1, rules = {}

r2:
itr1:
itr2:
cnt(fR1, fR3, fR4) = 1, rules = {fR1}
r2’[capital] = Beijing, cnt(fR3) = 1,
cnt(fR4) = 2, rules = {fR4}
r2’[city] = Shanghai, rules = {}

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	Shanghai	Hongkong	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

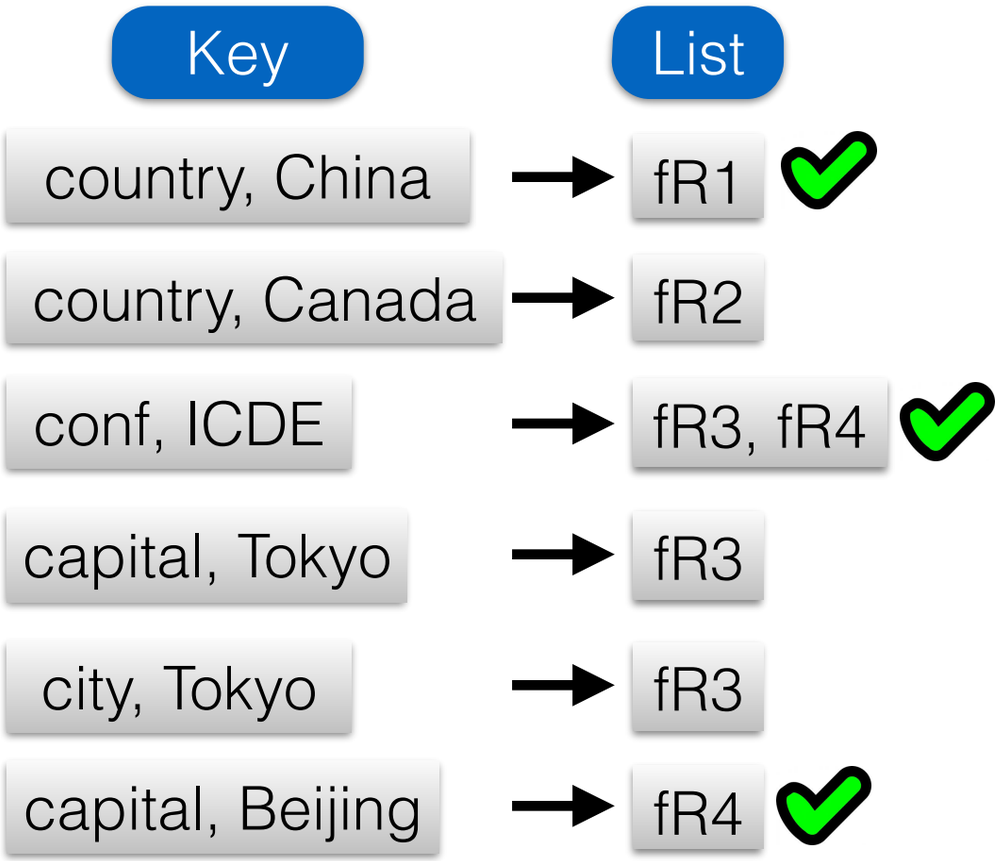
country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai



r1:
cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1}

itr1:
r1’ = r1, rules = {}

r2:
cnt(fR1, fR3, fR4) = 1, rules = {fR1}

itr1:
r2’[capital] = Beijing, cnt(fR3) = 1,
cnt(fR4) = 2, rules = {fR4}

itr2:
r2’[city] = Shanghai, rules = {}

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	**Shanghai**	ICDE
r3	Peter	China	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai Hongkong	Beijing

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Key		List	
country, China	→	fR1	✓
country, Canada	→	fR2	
conf, ICDE	→	fR3, fR4	✓
capital, Tokyo	→	fR3	
city, Tokyo	→	fR3	
capital, Beijing	→	fR4	✓

r1: itr1:	cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1} r1’ = r1, rules = {}
r2: itr1: itr2:	cnt(fR1, fR3, fR4) = 1, rules = {fR1} r2’[capital] = Beijing, cnt(fR3) = 1, cnt(fR4) = 2, rules = {fR4} r2’[city] = Shanghai, rules = {}

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	**Shanghai**	ICDE
r3	Peter	**Japan**	Tokyo	Tokyo	ICDE
r4	Mike	Canada	Toronto	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Key	List
country, China	fR1 ✓
country, Canada	fR2
conf, ICDE	fR3, fR4 ✓
capital, Tokyo	fR3
city, Tokyo	fR3
capital, Beijing	fR4 ✓

r1:
itr1:
cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1}
r1' = r1, rules = {}

r2:
itr1:
itr2:
cnt(fR1, fR3, fR4) = 1, rules = {fR1}
r2'[capital] = Beijing, cnt(fR3) = 1,
cnt(fR4) = 2, rules = {fR4}
r2'[city] = Shanghai, rules = {}

r3:
itr1:
cnt(fR3) = 3, cnt(fR4) = 1, rules = {fR3}
r3'[country] = Japan, rules = {}

Repairing with Fixing Rules

	name	country	capital	city	conf
r1	George	China	Beijing	Beijing	SIGMOD
r2	Ian	China	**Beijing**	**Shanghai**	ICDE
r3	Peter	**Japan**	Tokyo	Tokyo	ICDE
r4	Mike	Canada	**Ottawa**	Toronto	VLDB

fR_1

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

fR_2

country	{capital	capital
Canada	Toronto	Ottawa

fR_3

capital	city	conf	{country	country
Tokyo	Tokyo	ICDE	China	Japan

fR_4

capital	conf	{city	city
Beijing	ICDE	Hongkong	Shanghai

Key	List
country, China	fR1 ✓
country, Canada	fR2
conf, ICDE	fR3, fR4 ✓
capital, Tokyo	fR3
city, Tokyo	fR3
capital, Beijing	fR4 ✓

r1:
cnt(fR1) = 1, cnt(fR4) = 1, rules = {fR1}

itr1:
r1’ = r1, rules = {}

r2:
cnt(fR1, fR3, fR4) = 1, rules = {fR1}

itr1:
r2’[capital] = Beijing, cnt(fR3) = 1,
cnt(fR4) = 2, rules = {fR4}

itr2:
r2’[city] = Shanghai, rules = {}

r3:
cnt(fR3) = 3, cnt(fR4) = 1, rules = {fR3}

itr1:
r3’[country] = Japan, rules = {}

r4:
cnt(fR3) = 1, rules = {fR2}

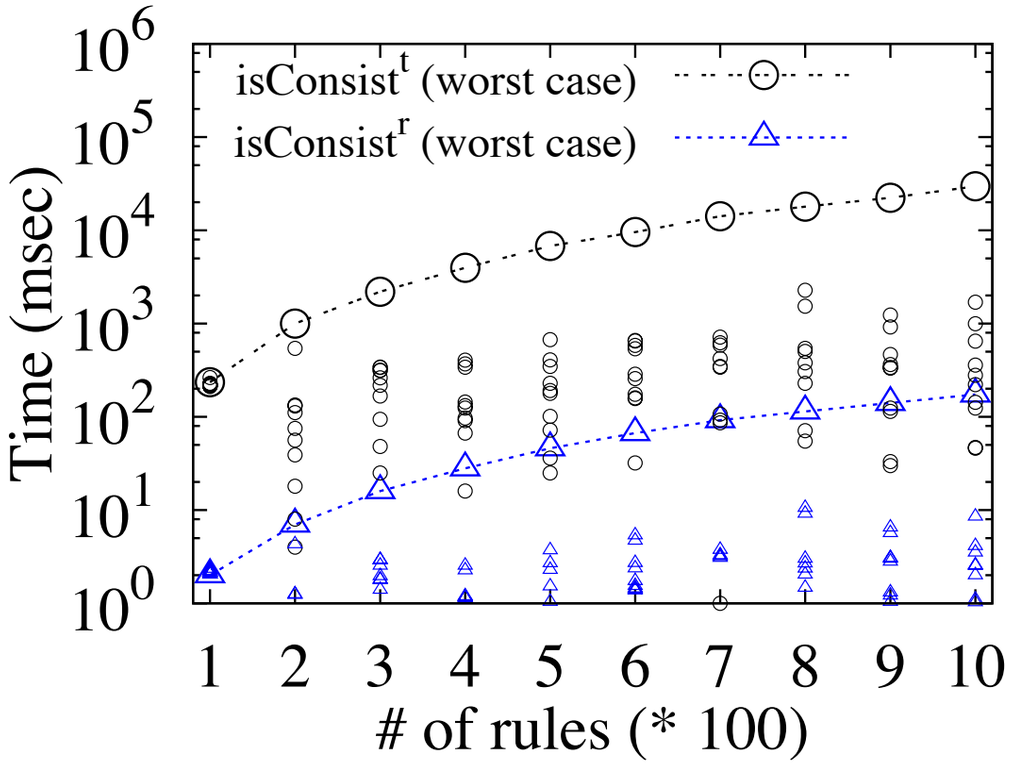
itr1:
r4’[capital] = Ottawa, rules = {}

Experiment

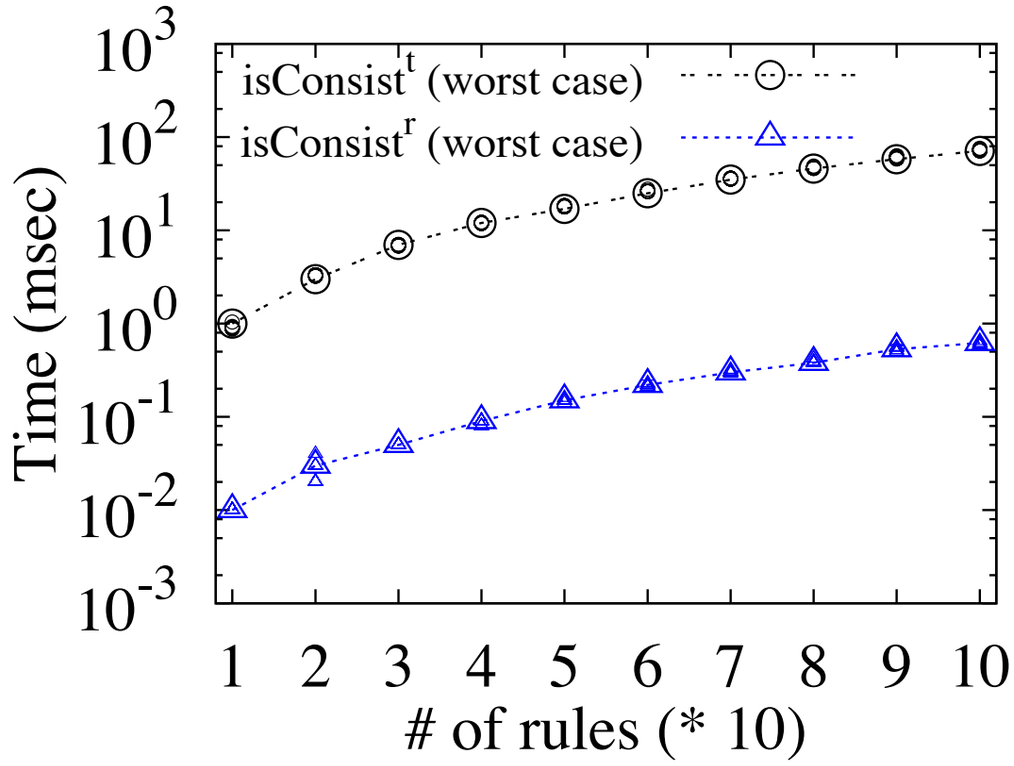
Experimental Study

- Efficiency of checking consistency
- Accuracy
- Efficiency of repairing algorithms

Efficiency of Checking Consistency

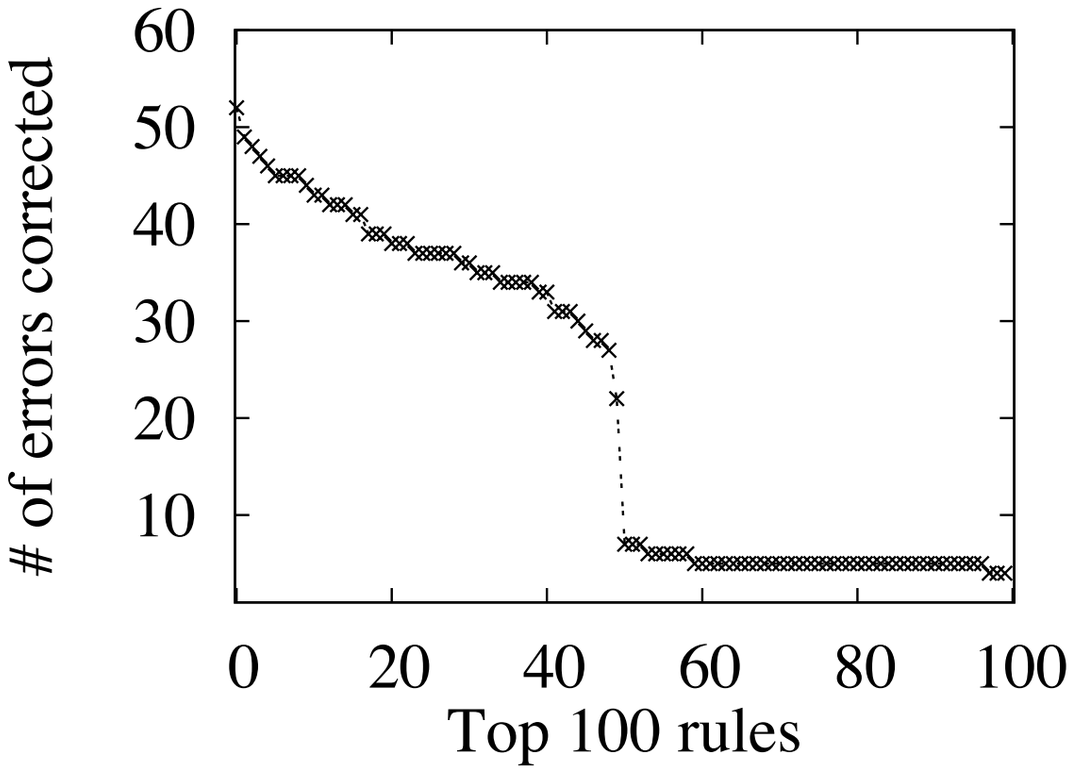


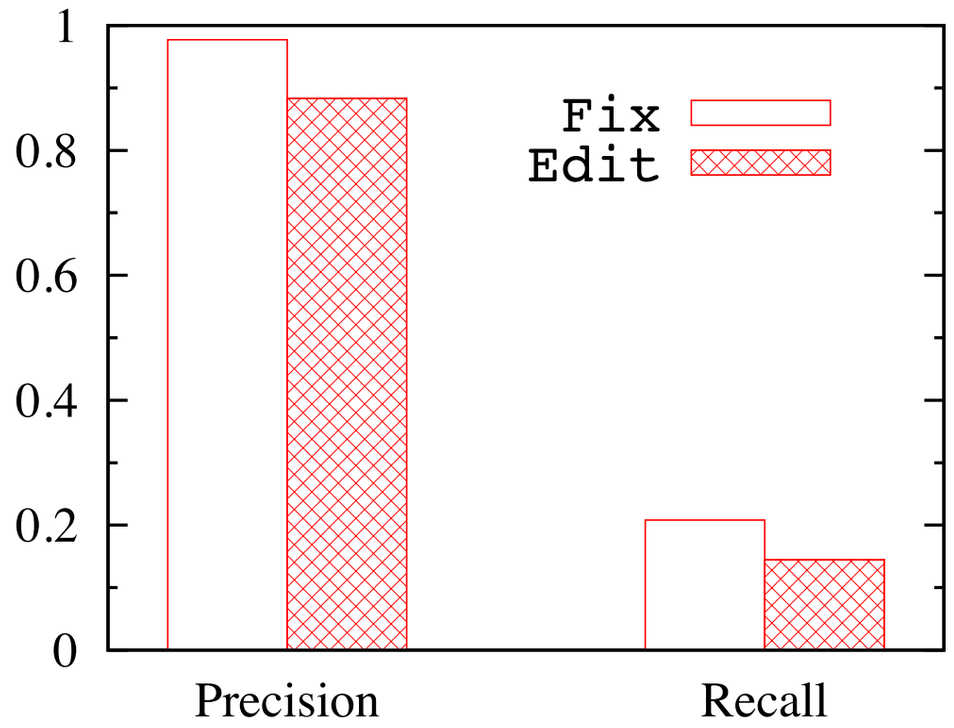
Hospital data



UIS

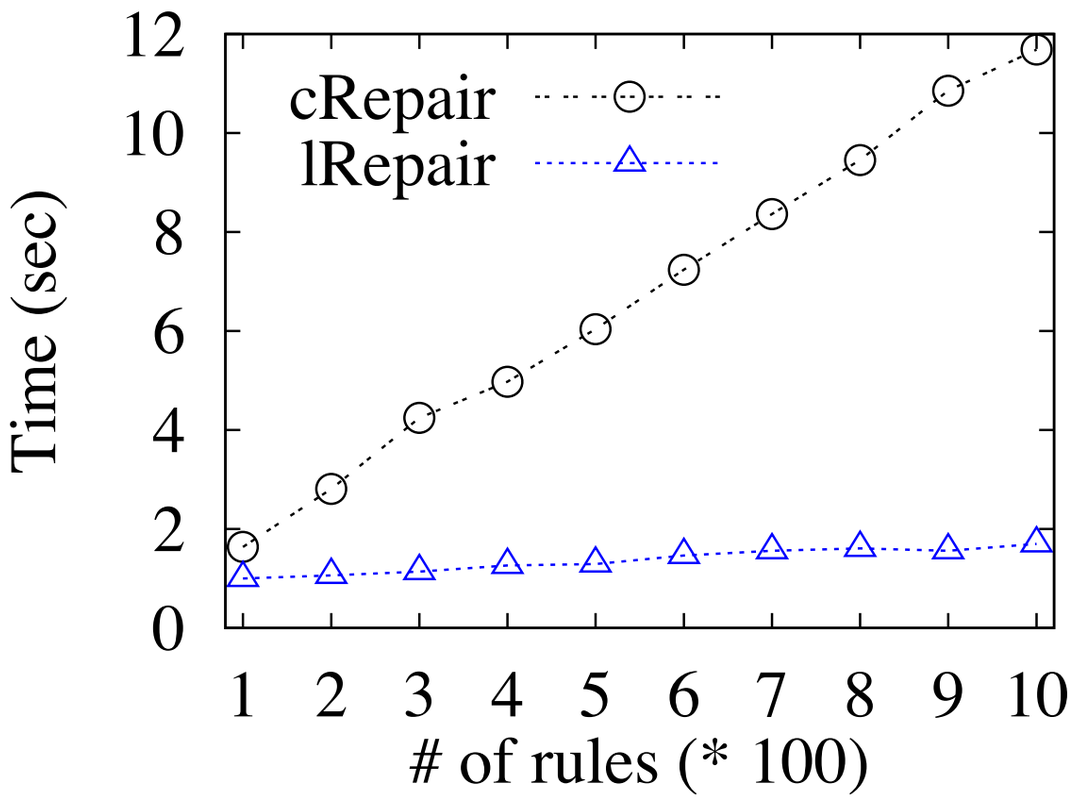
Accuracy



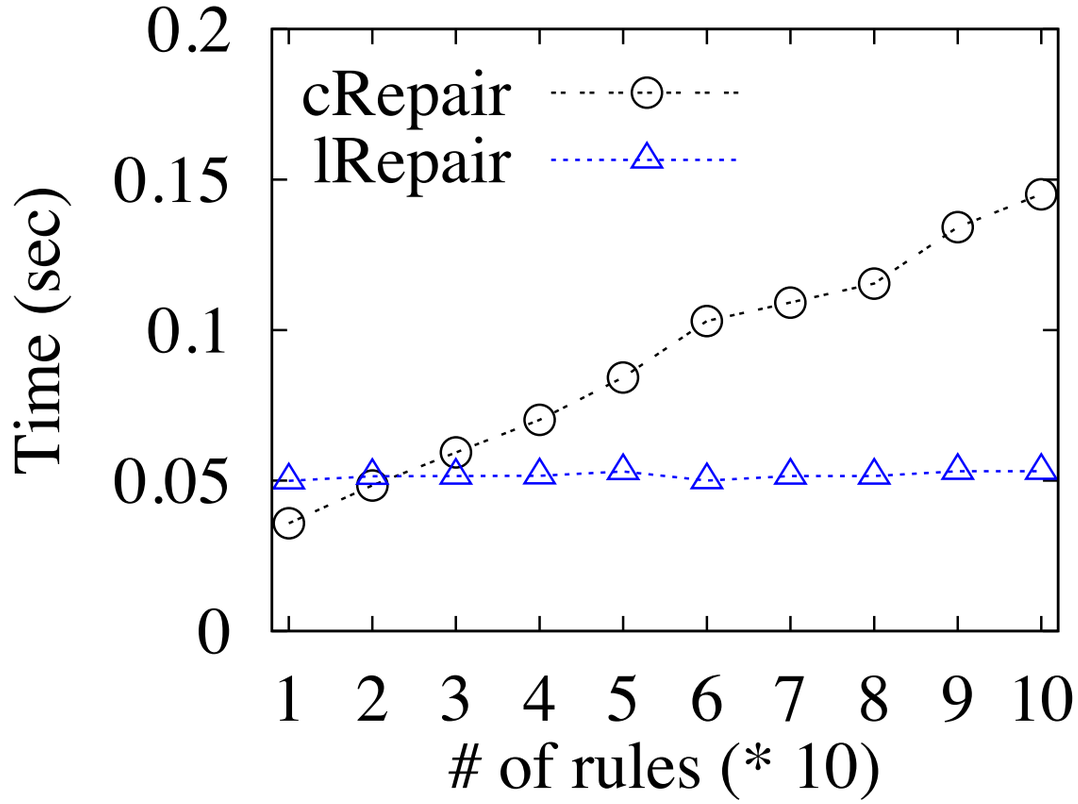


Hospital data

Efficiency of Repairing Algorithms



Hospital data



UIS

precision: +
recall: ++

Heuristic
(Automated)

precision: ++
recall: +

Fixing Rules

precision: ++
recall: ++

Certain
(User guided)

precision: +
recall: ++

precision: ++
recall: +

precision: ++
recall: ++

Conclusion:
Automated
Dependable
Fundamentals
Repair

Heuristic
(Automated)

Fixing Rules

Certain
(User guided)

precision: +
recall: ++

precision: ++
recall: +

precision: ++
recall: ++

Heuristic
(Automated)

Fixing Rules

Conclusion:
Automated
Dependable
Fundamentals
Repair

Future work:
Discovery
Generalized fixing rules

Certain
(User guided)

How to Get My Rules?

Generating Fixing Rules

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

Generating Fixing Rules

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

Freebase

MQL1

```
[{
    "type": "/location/country",
    "name": null,
    "/location/country/capital": []
}]
```

Generating Fixing Rules

country	{capital	capital
China	Shanghai	Beijing
	Hongkong	

Freebase

MQL1

```
[{
    "type": "/location/country",
    "name": null,
    "/location/country/capital": []
}]
```

MQL2

```
[{
    "/location/country/iso3166 1 shortname": "CHINA",
    "/location/location/contains": [{
        "name": null,
        "type": "/location/citytown"
    }]
}]
```