Data Cleaning

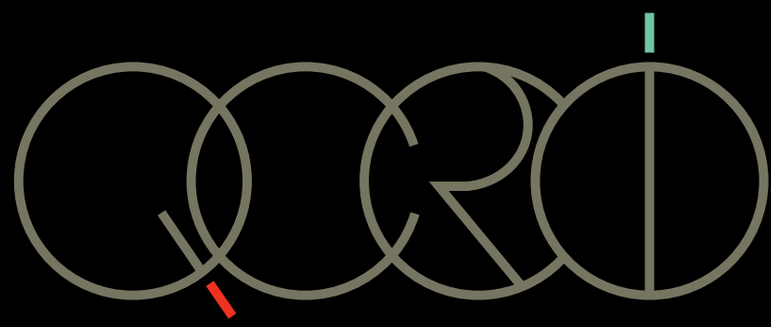
Nan Tang, QCRI

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Member of Qatar Foundation عضو في مؤسسة قطر

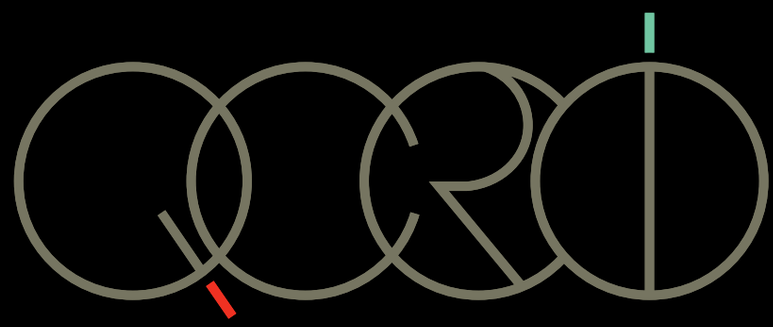
Big Data Cleaning

Nan Tang, QCRI



Big Data Cleaning

Nan Tang, QCRI



Data Cleaning?

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

Data Cleaning Market

INFORMATICA

Data Explorer

Tamr

Microsoft® SQL Server®

talend*
integration at any scale

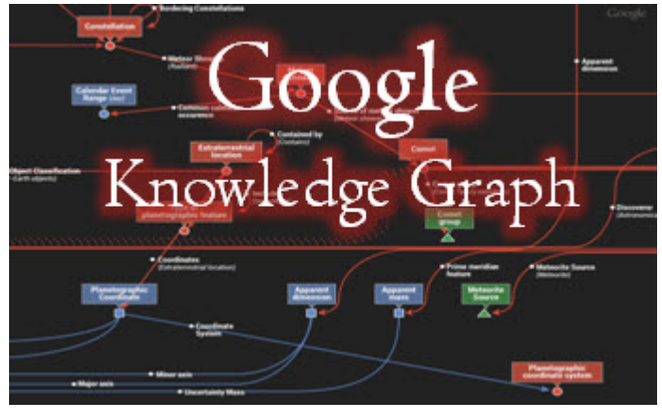
TRIFACTA

pentaho™
open source business intelligence™

ORACLE®
Data Quality 11g

IBM

InfoSphere™ software
Trusted Information



Data Cleaning Problems

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatari	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Data Cleaning Problems

typo

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatari	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Data Cleaning Problems

typo

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Data Cleaning Problems

typo

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Data Cleaning Problems

typo

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

Data Cleaning Problems

typo

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

Data Cleaning Problems

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

truth discovery

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

name	full
NTU	Nanyang Technological University
NUS	National University of Singapore

truth discovery

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

truth discovery

name	full
NTU	Nanyang Technological University
NUS	National University of Singapore

ETL (transformation)

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

truth discovery

name	full
NTU	Nanyang Technological University
NUS	National University of Singapore

ETL (transformation)

(UK, hasCapital, London)
KBs (e.g., Yago)

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

name	graduated	affiliation	country	capital	age
Nan Tang	CUHK	QCRI	Qatar	Doha	33
Xiaokui Xiao	CUHK	NTU	Singapore	Singapore	
Nan Tang	CUHK	University of Edinburgh	UK	Edinburgh	31
Gao Cong	NUS	University of Edinburgh	UK	London	36

Completeness

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

truth discovery

name	full
NTU	Nanyang Technological University
NUS	National University of Singapore

ETL (transformation)

Heterogeneous sources

(UK, hasCapital, London)
KBs (e.g., Yago)

Data Cleaning Problems

Duplicates

typo

Consistency

Currency

n	aduated	affiliation	country	capital	age
	HK	QCRI	Qatar	Doha	33
	HK	NTU	Singapore	Singapore	
Nan Tang	CUHK	Edinburgh	UK	Edinburgh	31
Gao Cong	NU	burgh	UK	London	36

Volume

Velocity

Completeness

V...

Heterogeneous sources

source2

name	affiliation
Nan Tang	QCRI

source3

name	affiliation
Nan Tang	CWI

... ...

name	
NTU	Na... Technological University
NUS	National University of Singapore

(UK, hasCapital, London)
KBs (e.g., Yago)

truth discovery

ETL (transformation)

Data Cleaning Solutions

Data Cleaning Solutions

CFD
(*ICDE 2012*)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Error detection

Fixing rules
(SIGMOD 2014)

... ...

Data Cleaning Solutions

CFD
(*ICDE 2012*)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

Error detection

input

Equivalence class

Set cover

Sat solver

Heuristic methods

Confidence values
(SIGMOD 2011)

Data repairing

User guided
(VLDB 2010 best paper)

Data Cleaning Solutions

CFD
(*ICDE 2012*)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014, ICDE2014 demo)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

input

Equivalence class

Set cover

Sat solver

Heuristic methods

Confidence values
(SIGMOD 2011)

Data repairing

User guided
(VLDB 2010 best paper)

Data Cleaning Solutions

CFD
(*ICDE 2012*)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014, ICDE2014 demo)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

input

Equivalence class

Set cover

Sat solver

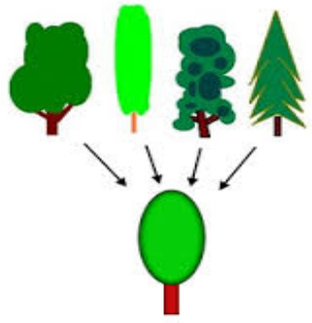
Heuristic methods

Confidence values
(SIGMOD 2011)

Data repairing

User guided
(VLDB 2010 best paper)

Generic system



NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

ER
(SIGMOD 2014 demo)

Open source

Data Cleaning Solutions

CFD
(*ICDE 2012*)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014, ICDE2014 demo)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

input

Equivalence class

Set cover

Sat solver

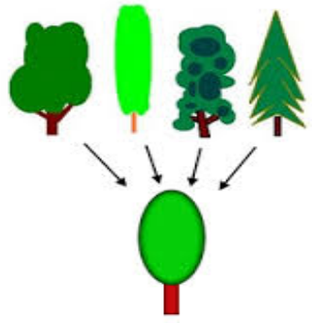
Heuristic methods

Confidence values
(SIGMOD 2011)

Data repairing

User guided
(VLDB 2010 best paper)

Generic system



NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

ER
(SIGMOD 2014 demo)

Open source

Error Detection

Error Detection

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

Error Detection

FD: [country] -> [capital]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

Error Detection

FD: [country] -> [capital]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

DC: ¬ t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

DC: ¬ t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

	country	capital
s1	China	Beijing
s2	Canada	Ottawa
s3	...	...

cap

DC: ¬ t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

	name	country	capital	city	salary	tax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yin	China	Shanghai	Hongkong	40000	1200
r3	Si	Netherlands	Den Hagg	Utrecht	60000	1400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

	country	capital
s1	China	Beijing
s2	Canada	Ottawa
s3	...	...

cap

DC: ¬ t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

MD: (emp[country] = cap[country]) -> (emp[capital] <=> cap[capital])

Error Detection

FD: [country] -> [capital]

CFD: [country = China] -> [capital = Beijing]

Inclusion dependency

Currency constraint

Sequential dependency

Aggregation constraint

Accuracy constraint

… … … …

			apita		ax
r1	Nan	China	Beijing	Beijing	50000	1000
r2	Yi	China	Shangh			200
r3			n Hag			400
r4	Lei	Netherlands	Amsterdam	Amsterdam	35000	800

emp

s1	China	Beijing
s2	Canada	Ottawa

cap

DC: ¬ t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

MD: (emp[country] = cap[country]) -> (emp[capital] <=> cap[capital])

Data Repairing

CFD
(ICDE 2012)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

input

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

Equivalence class
Set cover

Sat solver

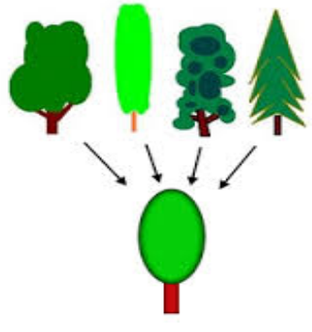
Heuristic methods

Confidence values
(SIGMOD 2011)

User guided
(VLDB 2010 best paper)

Data repairing

Generic system



NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

ER
(SIGMOD 2014 demo)

Open source
Commercialize

Data Repairing

CFD
(ICDE 2012)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

input

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

Equivalence class
Set cover

Sat solver

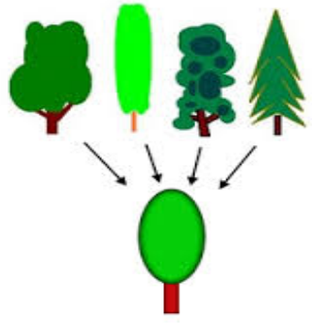
Heuristic methods

Confidence values
(SIGMOD 2011)

User guided
(VLDB 2010 best paper)

Data repairing

Generic system



NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

ER
(SIGMOD 2014 demo)

Open source
Commercialize

Automated

Computing a Consistent Database

D

Computing a Consistent Database

D

D_g ?

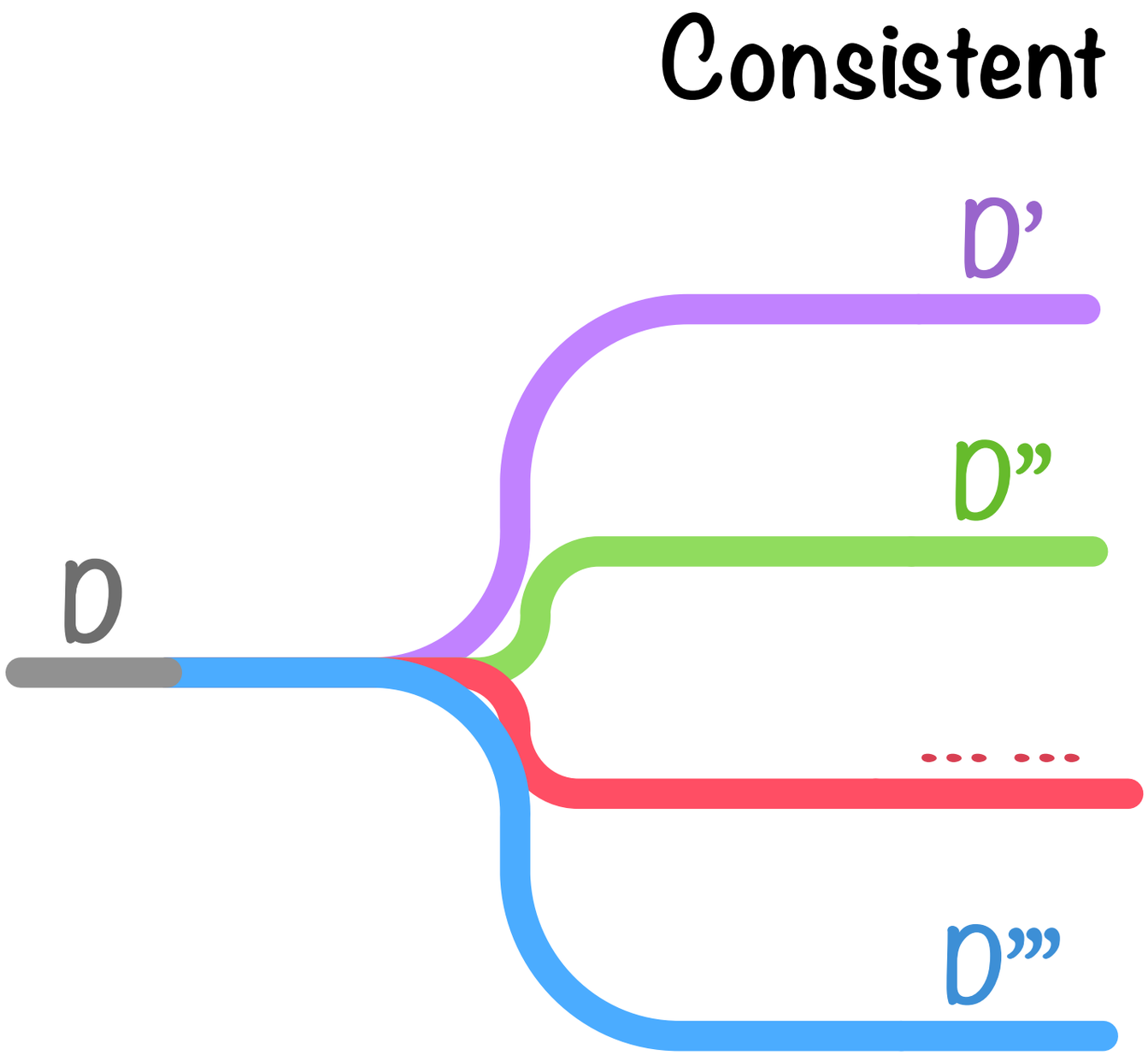
Computing a Consistent Database

Consistent

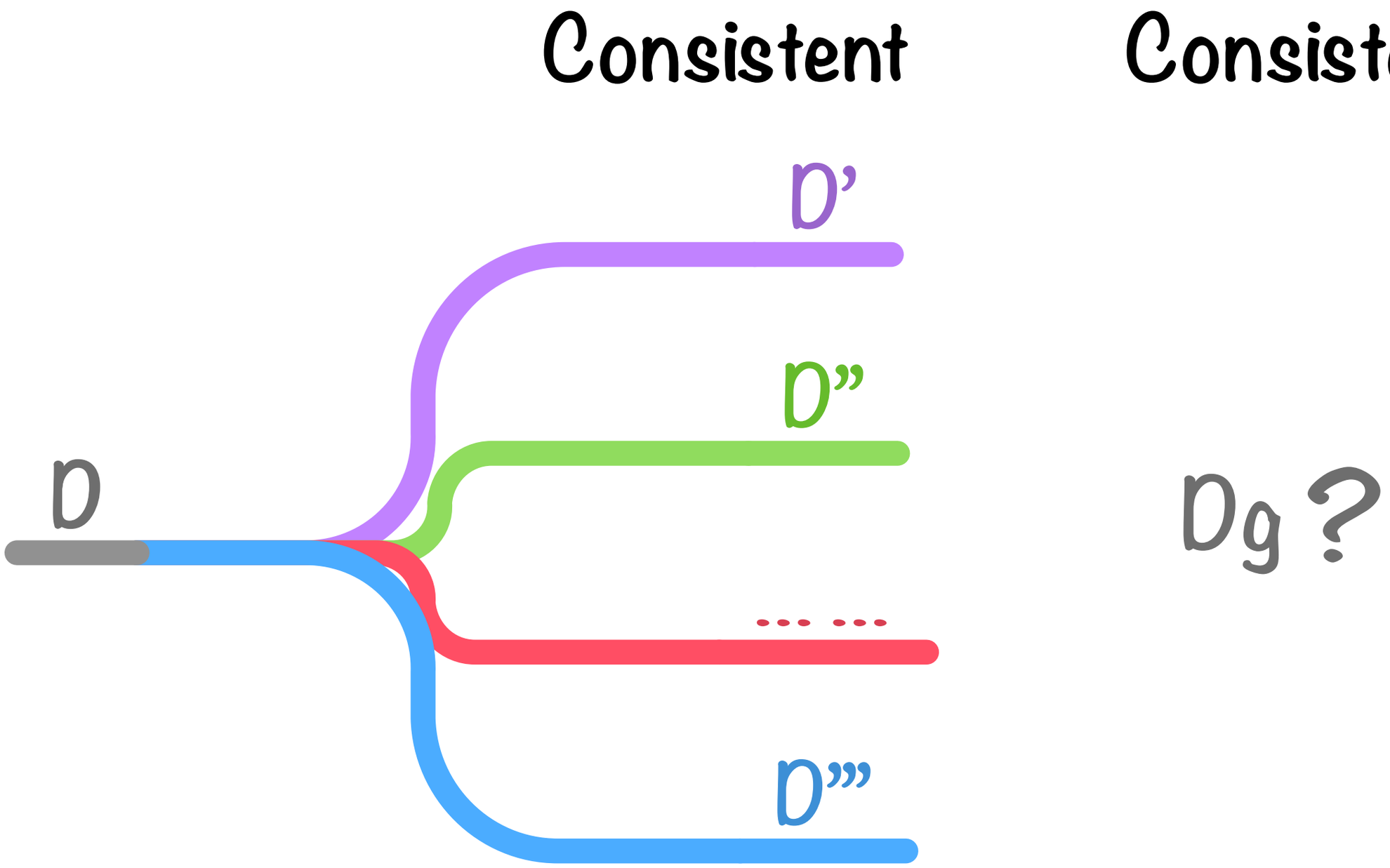
D

D_g ?

Computing a Consistent Database



Computing a Consistent Database



find a D' such that dist(D,D') is minimum

Computing a Consistent Database

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo	3	Kyoto	35000	800

Computing a Consistent Database

FD1: [nationality] -> [capital]

FD2: [areacode] -> [capital]

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo	3	Kyoto	35000	800

Computing a Consistent Database

FD1: [nationality] -> [capital]

FD2: [areacode] -> [capital]

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo	3	Kyoto	35000	800

Computing a Consistent Database

FD1: [nationality] -> [capital]

FD2: [areacode] -> [capital]

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai **Beijing**	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo **Beijing**	3	Kyoto	35000	800

Computing a Consistent Database

FD1: [nationality] -> [capital]

FD2: [areacode] -> [capital]

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai **Beijing**	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo **Beijing**	3	Kyoto	35000	800

Equivalence class

Vertex cover

SAT solver

… …

Computing a Consistent Database

FD1: [nationality] -> [capital]

FD2: [areacode] -> [capital]

	name	nationality	capital	areacode	bornAt	salary	tax
r1	Nan	China	Beijing	10	Shenyang	50000	1000
r2	Yan	China	Shanghai **Beijing**	10	Hangzhou	40000	900
r3	Si	China	Beijing	10	Changsha	60000	1400
r4	Miura	China	Tokyo **Beijing**	3	Kyoto	35000	800

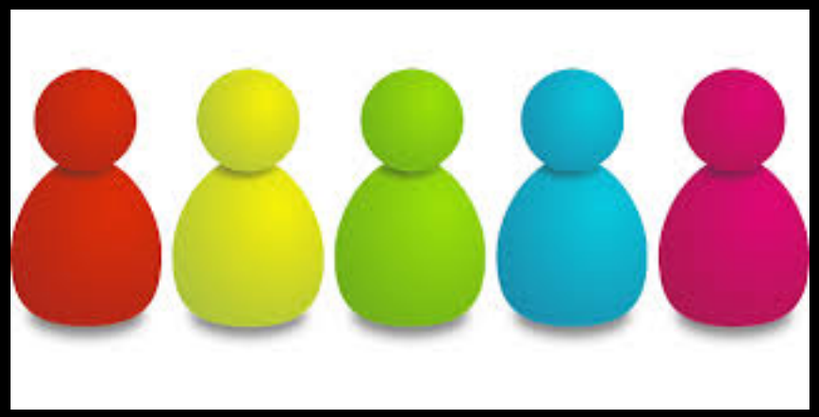
Equivalence class

Vertex cover

SAT solver

… …

User Guided



Certain Fixes (VLDB 2010 Best Paper)

	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

Certain Fixes (VLDB 2010 Best Paper)

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

Certain Fixes (VLDB 2010 Best Paper)

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

Certain Fixes (VLDB 2010 Best Paper)

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.

Certain Fixes (VLDB 2010 Best Paper)

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.

Certain Fixes (VLDB 2010 Best Paper)

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?

… … … …

Certain Fixes (VLDB 2010 Best Paper)

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)

Data patterns

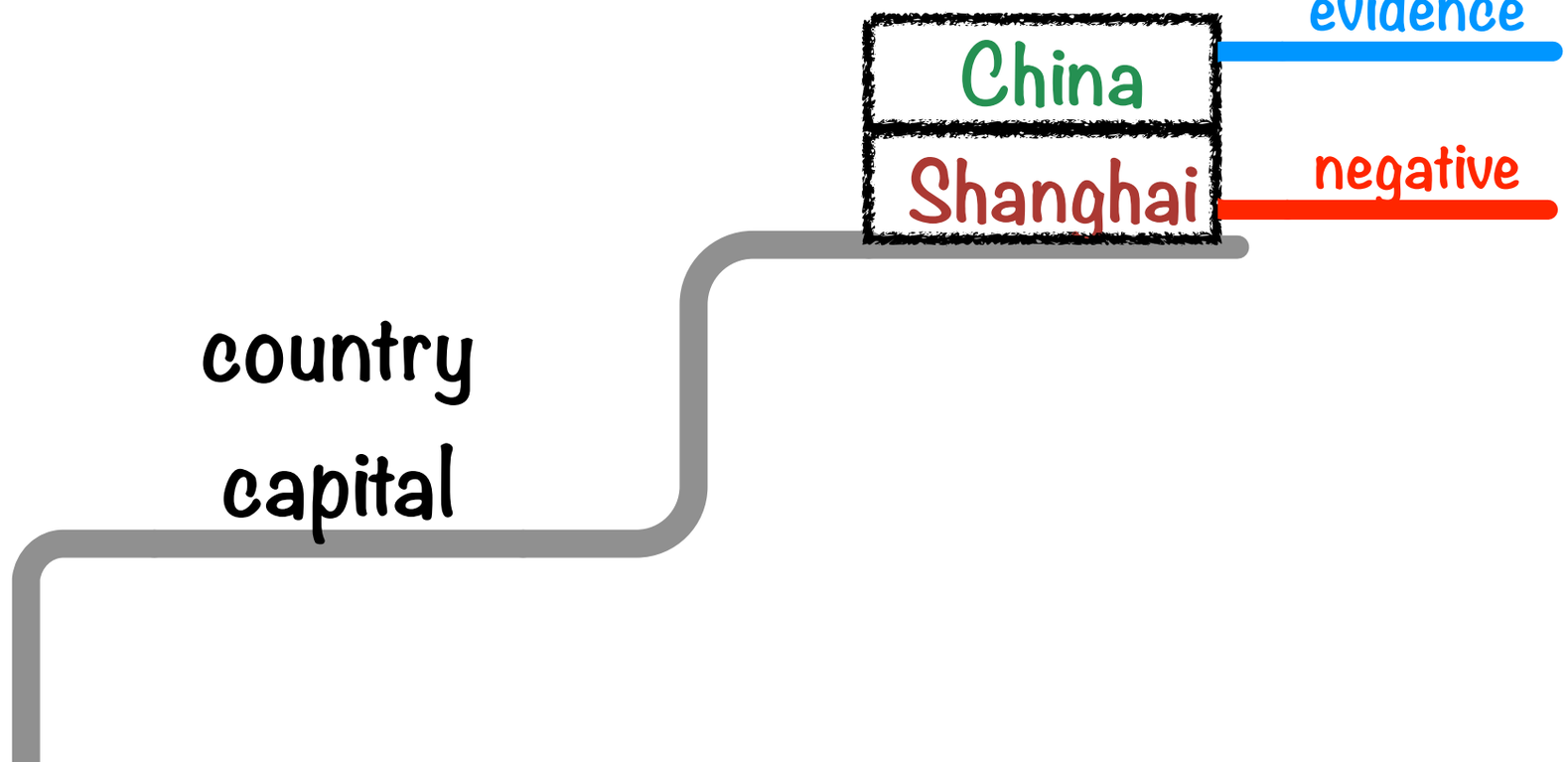
country

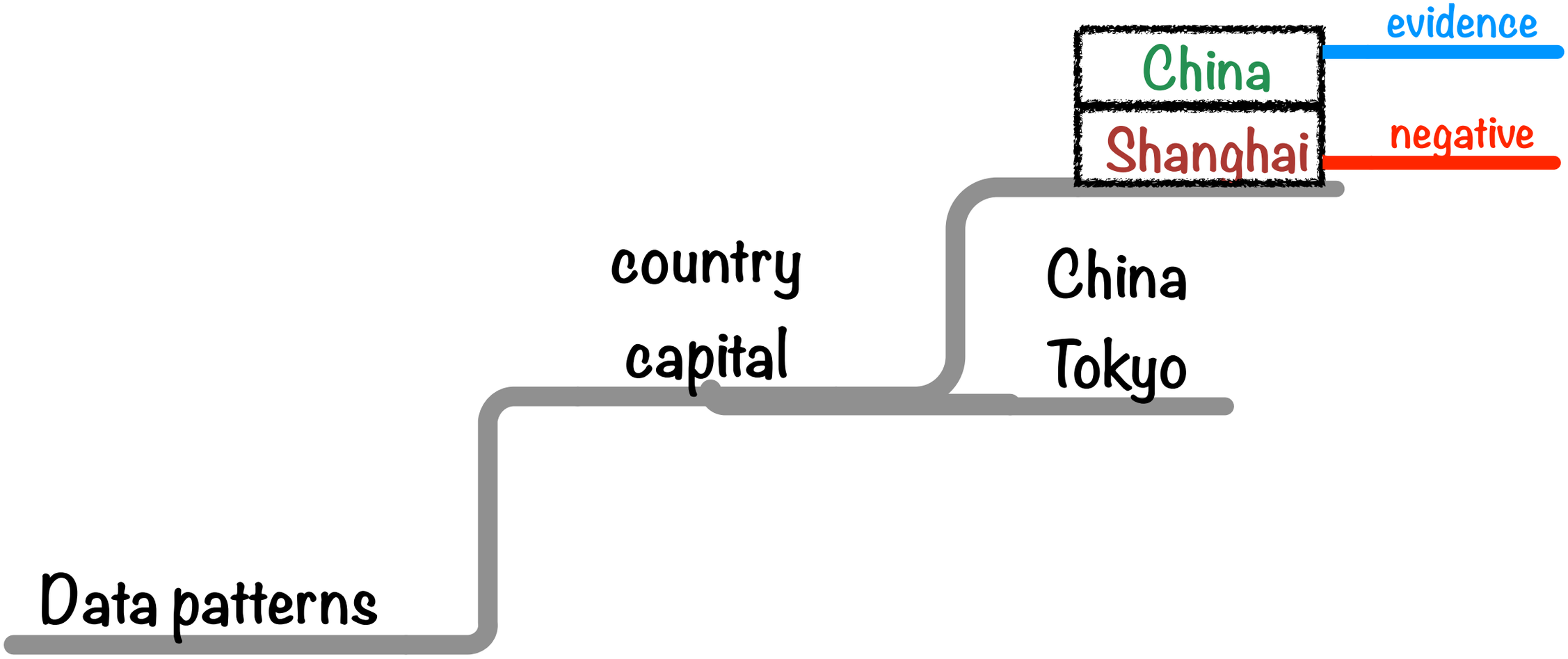
capital

China

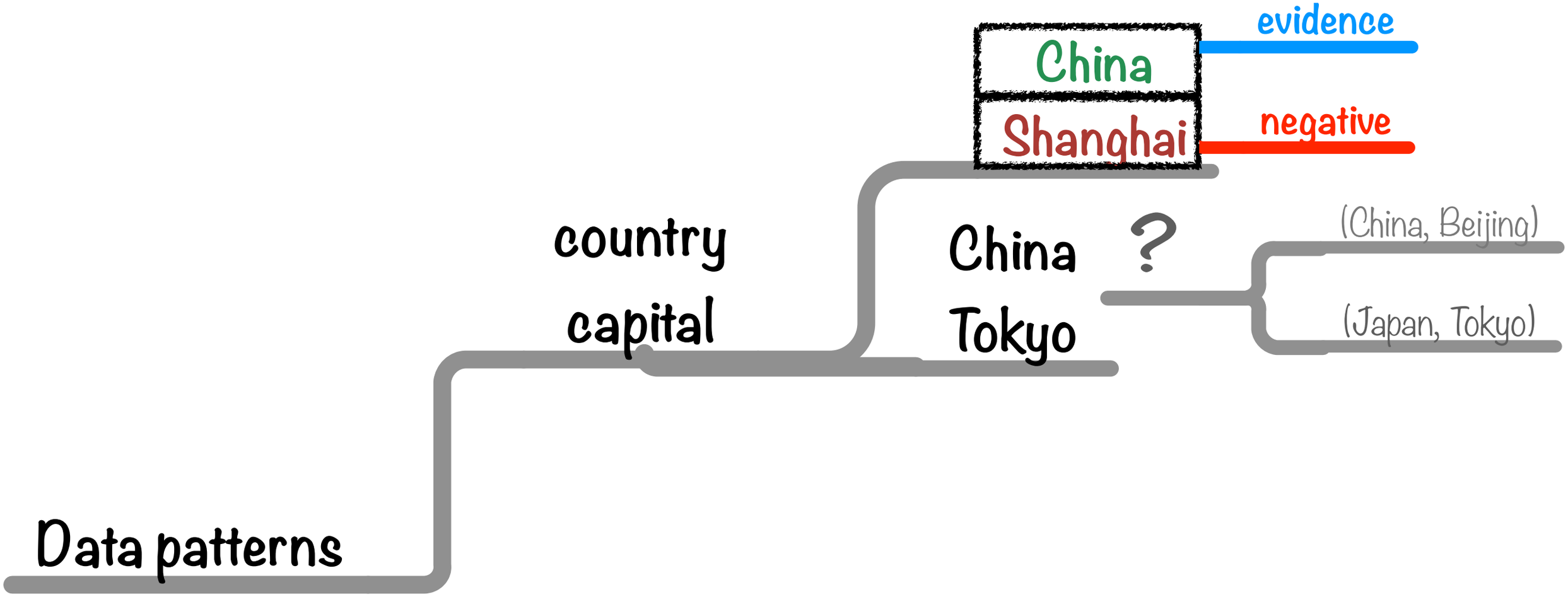
Shanghai

Data patterns

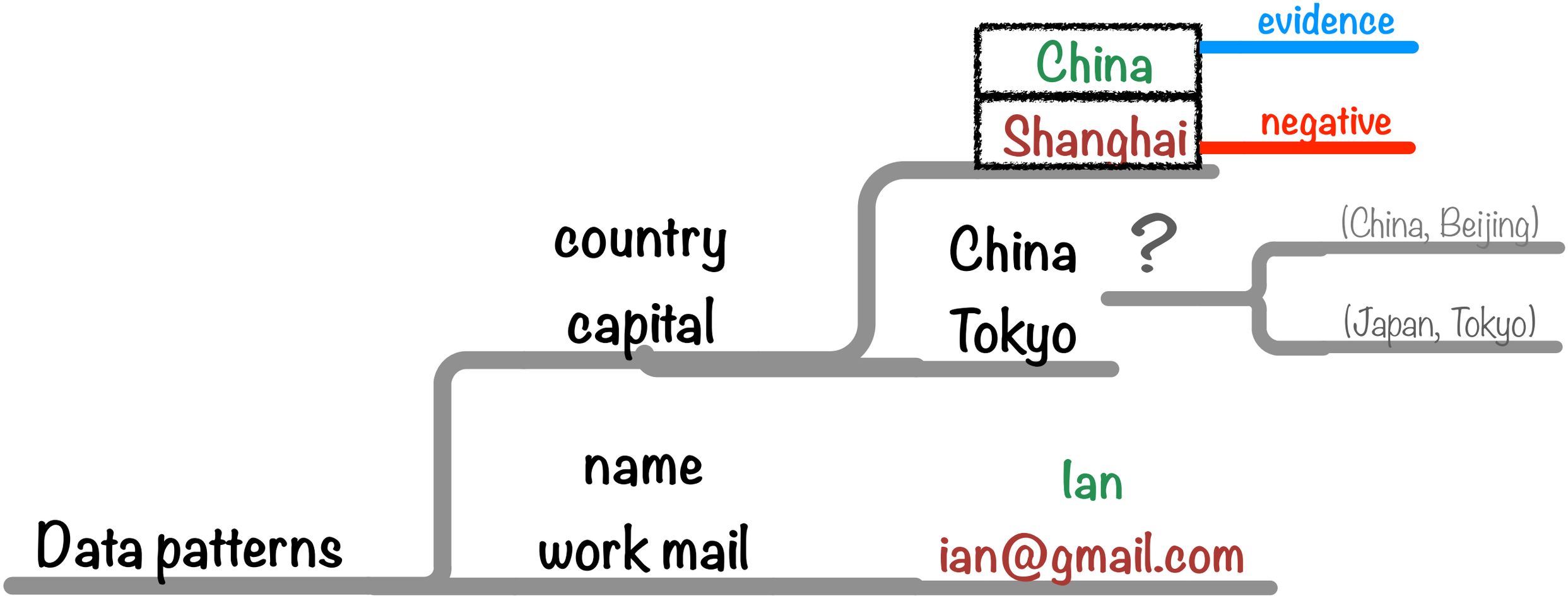




Data patterns



Data patterns


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

country

capital

China

Shanghai

evidence

negative

China ?

Tokyo

(China, Beijing)

(Japan, Tokyo)

name

work mail

Ian

ian@gmail.com

evidence

negative

Data patterns

Data patterns

country
capital

China
Tokyo

China ?

China
Shanghai

evidence

negative

(China, Beijing)

(Japan, Tokyo)

name

work mail

Ian
ian@gmail.com

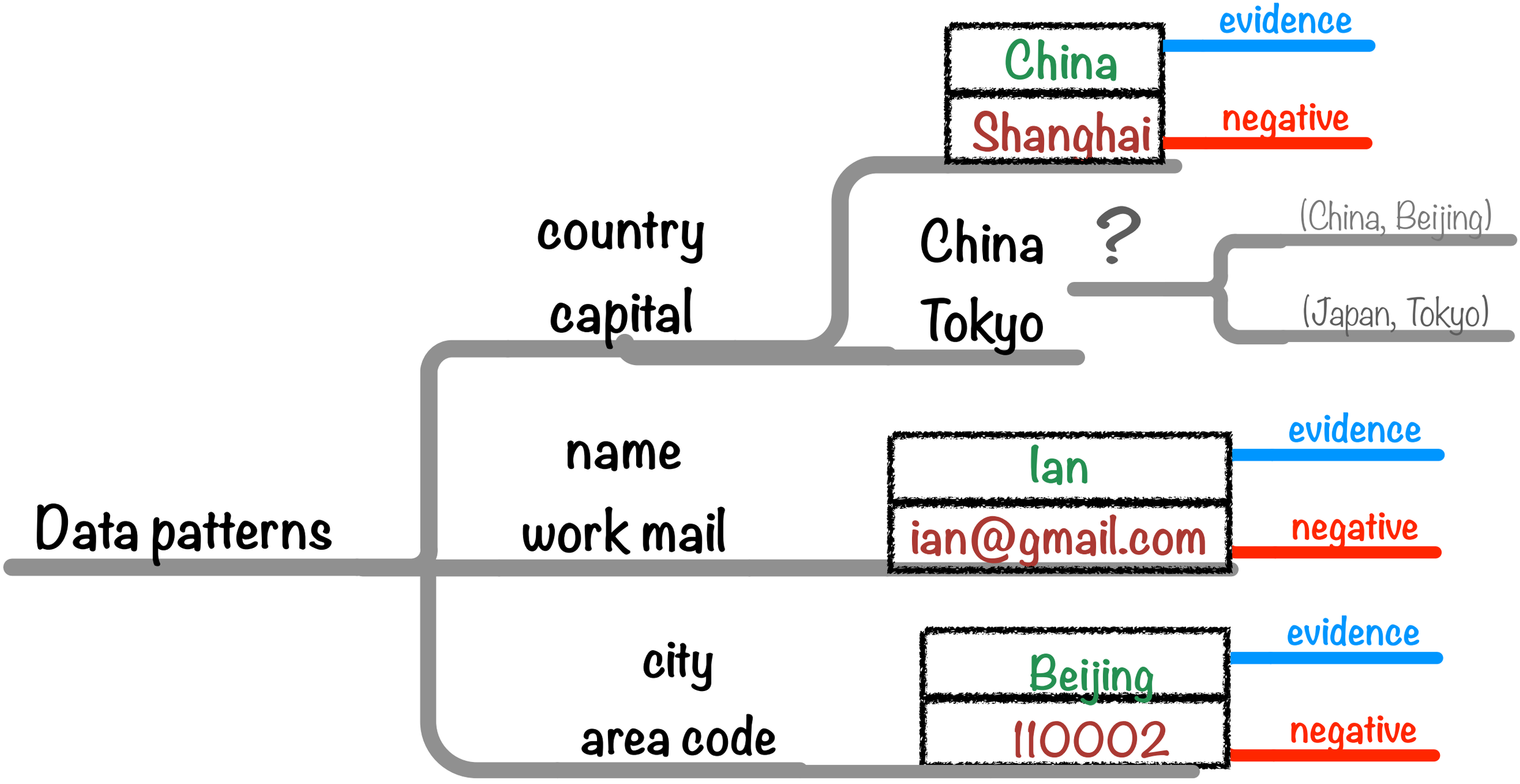
evidence

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 (SIGMOD 2014)

- **Syntax**

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

Fixing Rules (SIGMOD 2014)

- **Syntax**

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

country	{capital	capital
China	Shanghai Hongkong	Beijing

Fixing Rules (SIGMOD 2014)

- **Syntax**

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

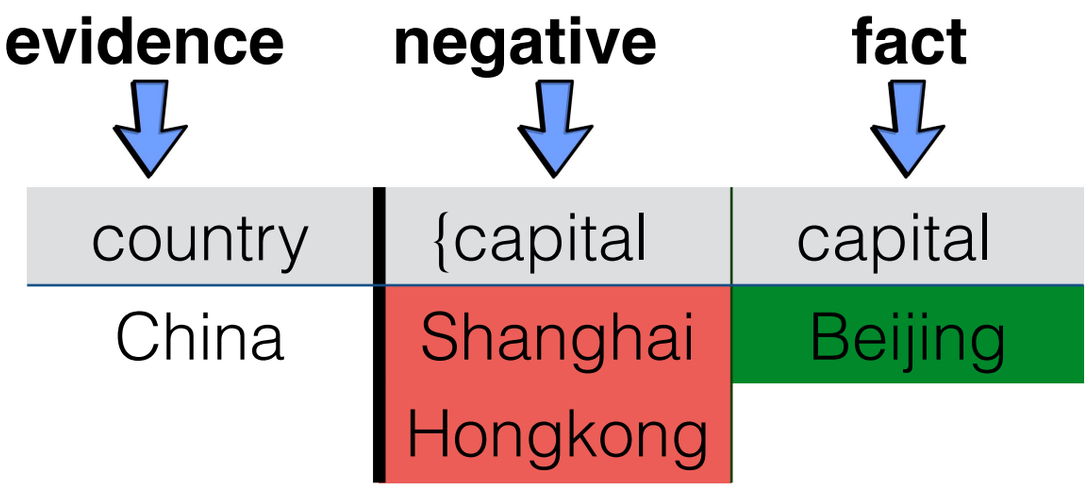
evidence **negative**

country	{capital	capital
China	Shanghai Hongkong	Beijing

Fixing Rules (SIGMOD 2014)

- **Syntax**

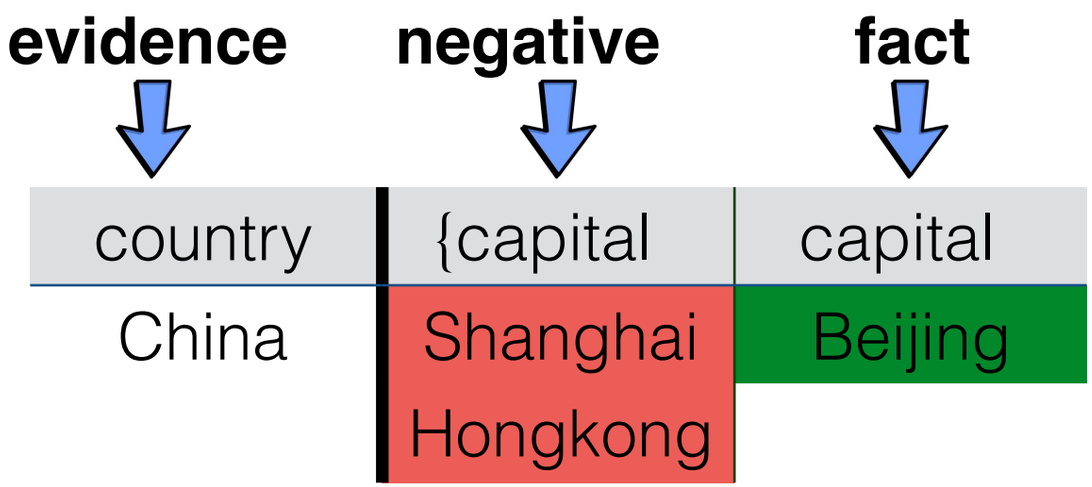
fR1: (([country], [China]), (capital, {Shanghai, Hongkong})) -> Beijing



Fixing Rules (SIGMOD 2014)

- **Syntax**

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

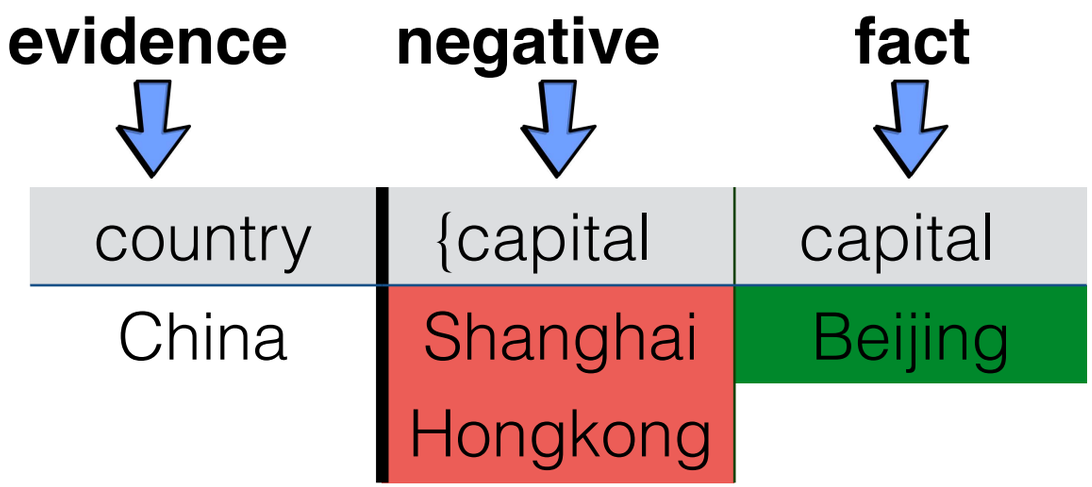


	name	nationality	capital	bornAt
r1	Nan	China	Beijing	Shenyang
r2	Yan	China	Shanghai	Hangzhou
r3	Si	China	Beijing	Changsha
r4	Miura	China	Tokyo	Kyoto

Fixing Rules (SIGMOD 2014)

- **Syntax**

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



	name	nationality	capital	bornAt
r1	Nan	China	Beijing	Shenyang
r2	Yan	China	Shanghai	Hangzhou
r3	Si	China	Beijing	Changsha
r4	Miura	China	Tokyo	Kyoto

Fixing Rules (SIGMOD 2014)

- **Syntax**

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

evidence **negative** **fact**

country	{capital	capital
China	Shanghai Hongkong	Beijing

	name	nationality	capital	bornAt
r1	Nan	China	Beijing	Shenyang
r2	Yan	China	Beijing	Hangzhou
r3	Si	China	Beijing	Changsha
r4	Miura	China	Tokyo	Kyoto

Fixing Rules (SIGMOD 2014)

- **Syntax**

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

evidence	negative	fact
country	{capital	capital
China	Shanghai Hongkong	Beijing

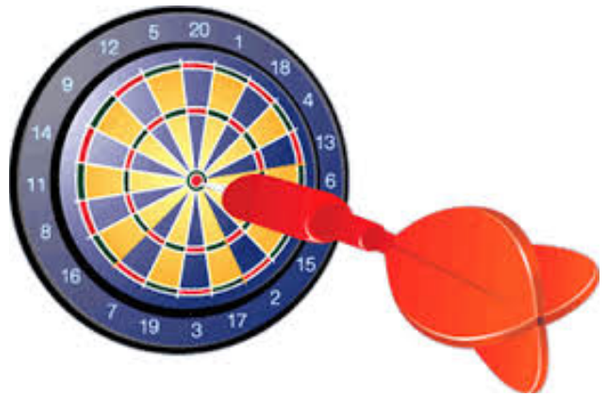
	name	nationality	capital	bornAt
r1	Nan	China	Beijing	Shenyang
r2	Yan	China	Beijing	Hangzhou
r3	Si	China	Beijing	Changsha
r4	Miura	China	Tokyo	Kyoto

Fixing Rules (SIGMOD 2014)

- **Syntax**

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

evidence	negative	fact
country	{capital	capital
China	Shanghai Hongkong	Beijing



	name	nationality	capital	bornAt
r1	Nan	China	Beijing	Shenyang
r2	Yan	China	Beijing	Hangzhou
r3	Si	China	Beijing	Changsha
r4	Miura	China	Tokyo	Kyoto

Confidence values Interaction

... ...

Matching and Repairing (SIGMOD 2011)

FD: [nationality] -> [capital]

MD: ((nationality, country) -> (capital, capital))

	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	Toronto (0.5)	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

	country	capital
s1	China (1.0)	Beijing (1.0)
s2	Canada (1.0)	Ottawa (1.0)
s3	Japan (1.0)	Tokyo (1.0)

Matching and Repairing (SIGMOD 2011)

FD: [nationality] -> [capital]

MD: ((nationality, country) -> (capital, capital))

	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	Toronto (0.5)	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

	country	capital
s1	China (1.0)	Beijing (1.0)
s2	Canada (1.0)	Ottawa (1.0)
s3	Japan (1.0)	Tokyo (1.0)

Matching and Repairing (SIGMOD 2011)

FD: [nationality] -> [capital]

MD: ((nationality, country) -> (capital, capital))

	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	Toronto (0.5)	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

	country	capital
s1	China (1.0)	Beijing (1.0)
s2	Canada (1.0)	Ottawa (1.0)
s3	Japan (1.0)	Tokyo (1.0)

Matching and Repairing (SIGMOD 2011)

FD: [nationality] -> [capital]

MD: ((nationality, country) -> (capital, capital))

	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	**Ottawa (1.0)**	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

	country	capital
s1	China (1.0)	Beijing (1.0)
s2	Canada (1.0)	Ottawa (1.0)
s3	Japan (1.0)	Tokyo (1.0)

Matching and Repairing (SIGMOD 2011)

FD: [nationality] -> [capital]

MD: ((nationality, country) -> (capital, capital))

	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	**Ottawa (1.0)**	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

	country	capital
s1	China (1.0)	Beijing (1.0)
s2	Canada (1.0)	Ottawa (1.0)
s3	Japan (1.0)	Tokyo (1.0)

Summary of Data Repairing

Consistent database

(heuristic)

Equivalence class
Vertex cover
Sat solver

Summary of Data Repairing

Consistent database (heuristic)

Equivalence class
Vertex cover
Sat solver

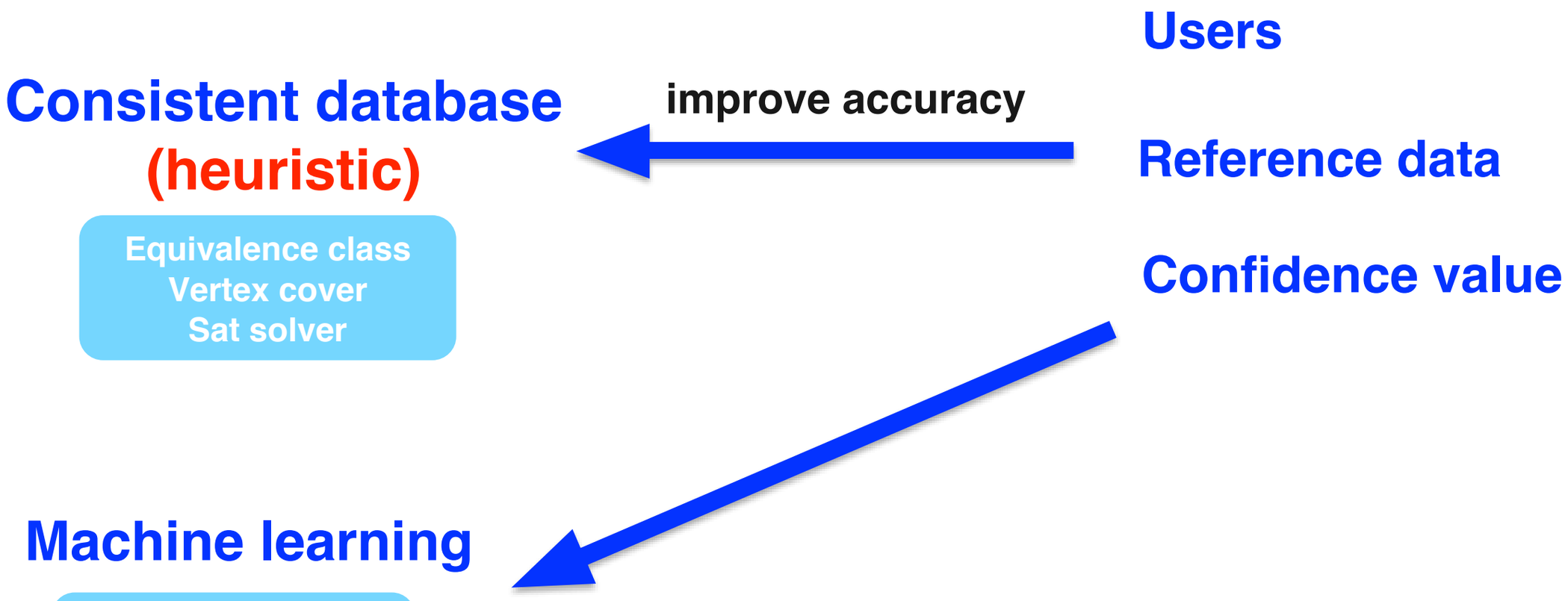
improve accuracy

Users

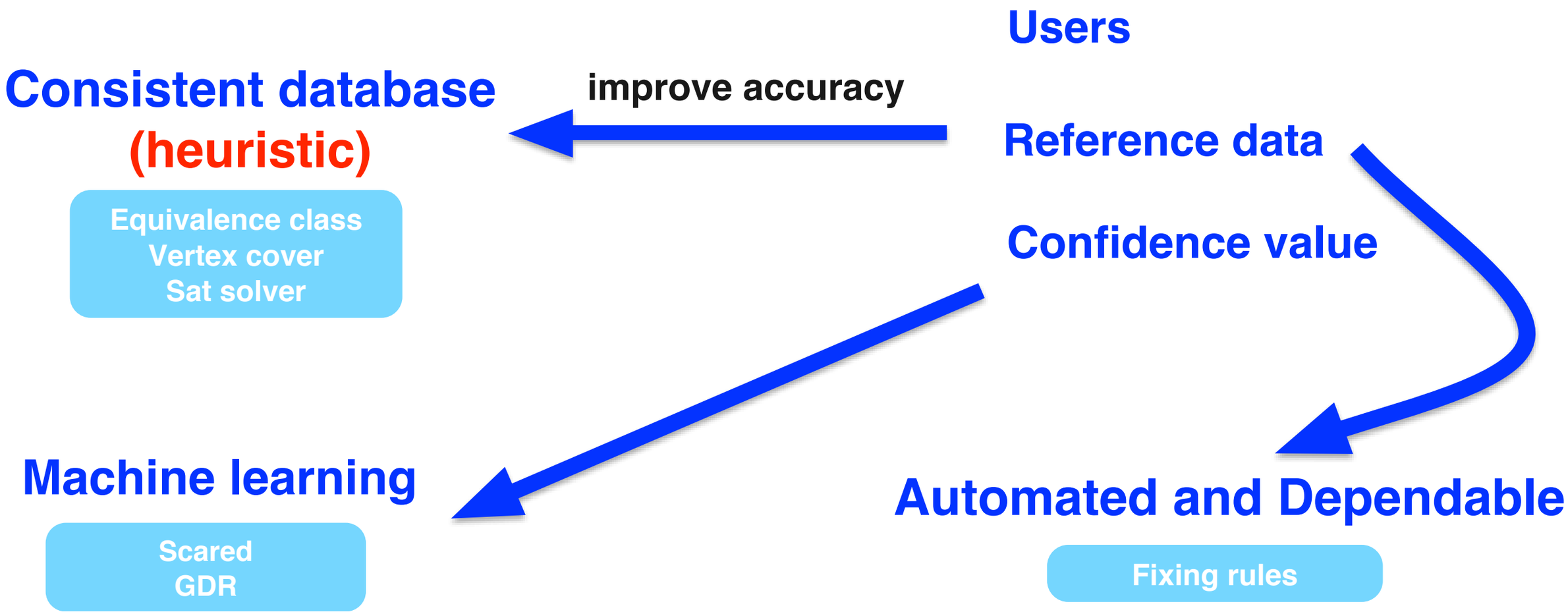
Reference data

Confidence value

Summary of Data Repairing



Summary of Data Repairing



Generic Data Cleaning System

CFD
(ICDE 2012)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

input

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

Equivalence class
Set cover

Sat solver

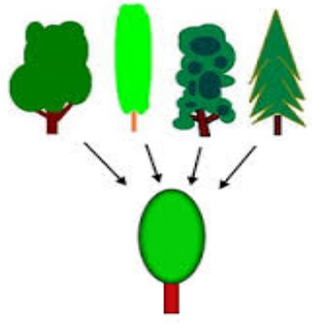
Heuristic methods

Confidence values
(SIGMOD 2011)

User guided
(VLDB 2010 best paper)

Data repairing

Generic system



NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

ER
(SIGMOD 2014 demo)

Open source
Commercialize

Generic Data Cleaning System

CFD
(ICDE 2012)

Functional dependency

Denial constraints

Currency constraints
(ICDE 2013)

Fixing rules
(SIGMOD 2014)

... ...

input

Error detection

input

Rule discovery

CFDs

DCs
(QCRI, VLDB 2014)

Unique columns
(QCRI, ICDE 2013, VLDB 2014)

Equivalence class
Set cover

Sat solver

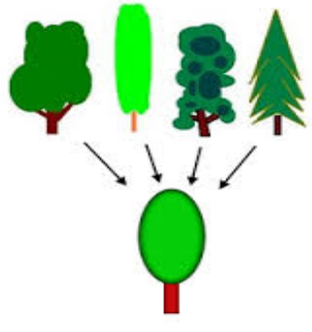
Heuristic methods

Confidence values
(SIGMOD 2011)

User guided
(VLDB 2010 best paper)

Data repairing

Generic system



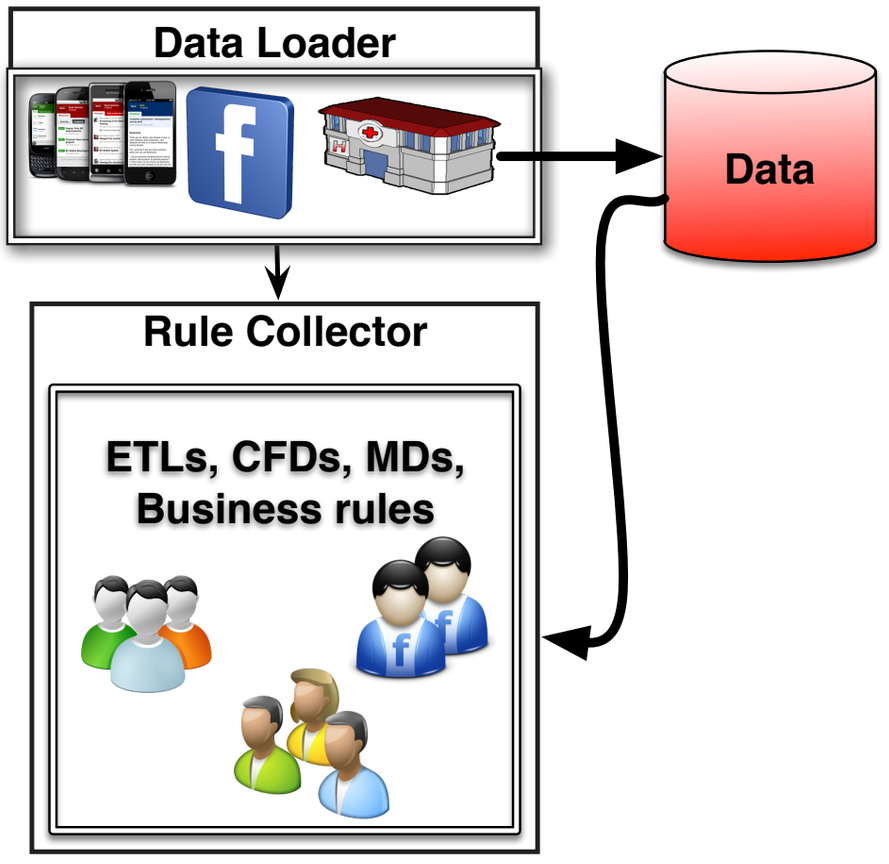
NADEEF
(SIGMOD 2013)

Dashboard
(VLDB 2013 demo)

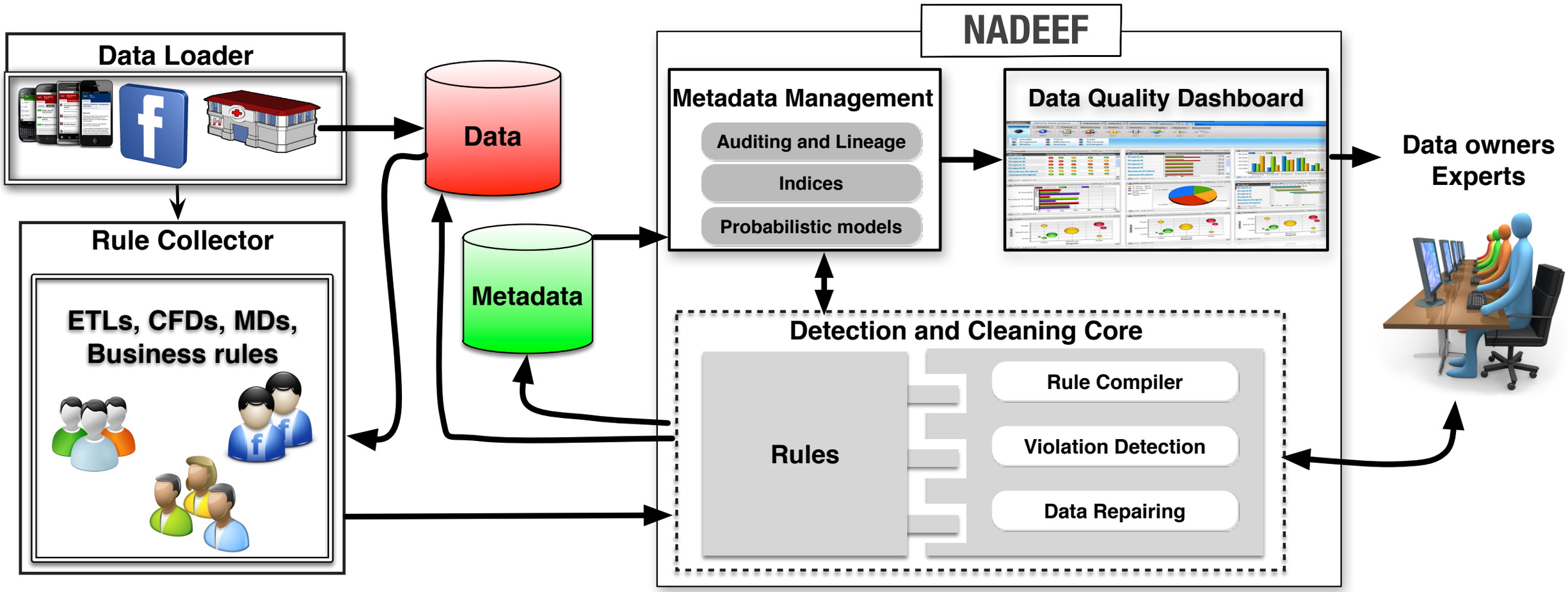
ER
(SIGMOD 2014 demo)

Open source
Commercialize

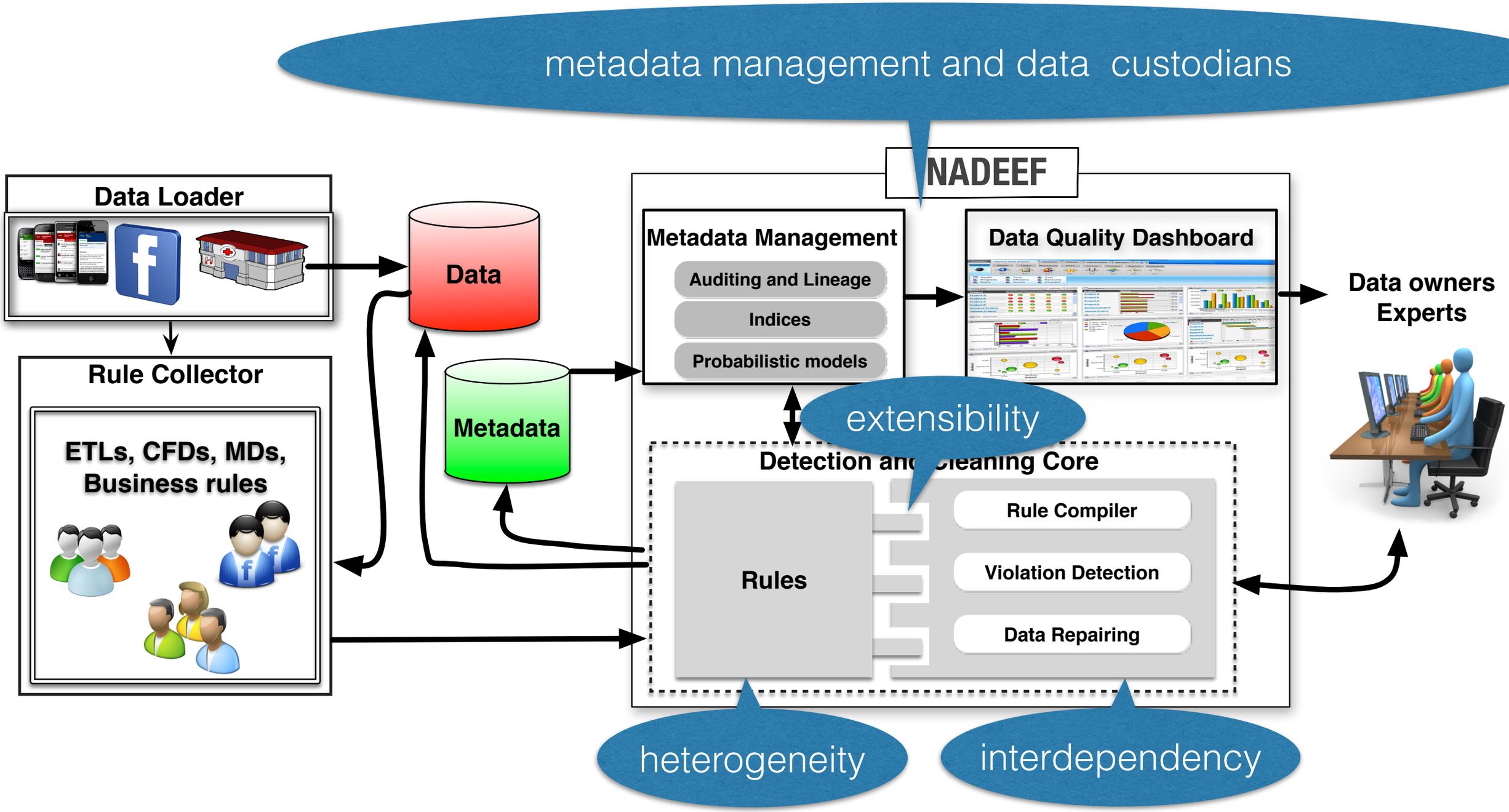
NADEEF (SIGMOD 2013)



NADEEF (SIGMOD 2013)



NADEEF (SIGMOD 2013)



NADEEF (SIGMOD 2013)

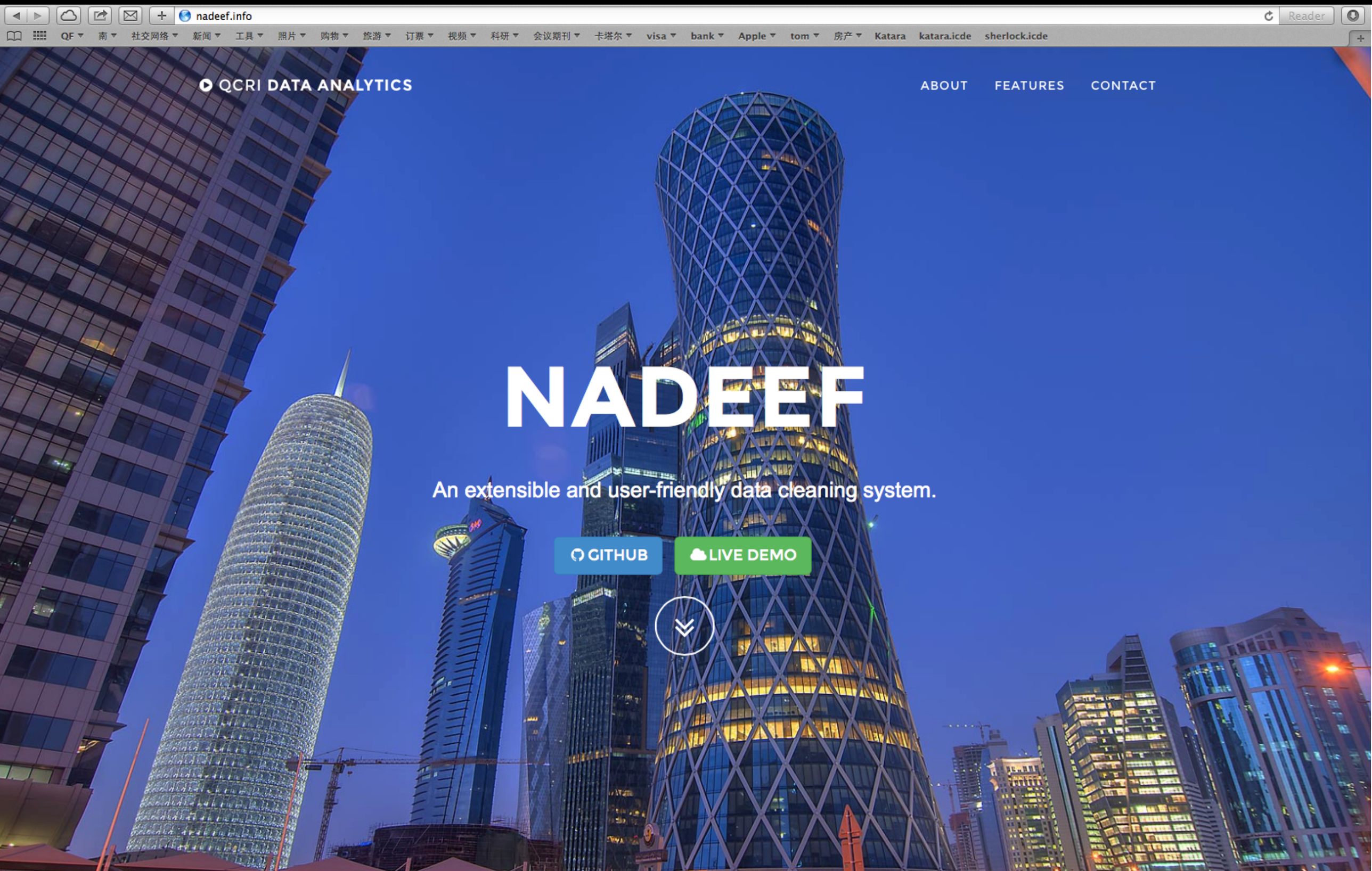
Rule Editor

- Detect
- Repair
- Block
- Iterator

```
 8     @Override
 9     public Collection<Violation> detect(TuplePair tuplePair) {
10         List<Violation> result = new ArrayList<>();
11         Tuple left = tuplePair.getLeft();
12         Tuple right = tuplePair.getRight();
13
14         if (
15             Metrics.getEqual(
16                 left.get("name"), right.get("name")) == 1.0 &&
17             Metrics.getLevenshtein(
18                 left.get("address"), right.get("address")) > 0.8 &&
19             Metrics.getEqual(
20                 left.get("gender"), right.get("gender")) == 1.0
21         ) {
22             Violation v = new Violation(getRuleName());
23             v.addTuple(left);
24             v.addTuple(right);
25             result.add(v);
26         }
27         return result;
28     }
29
30
```

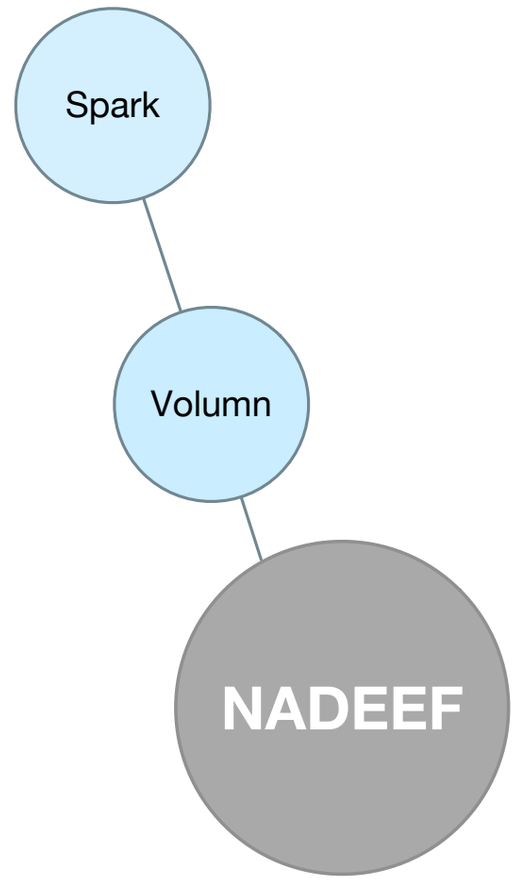
Close | Save changes

NADEEF Online

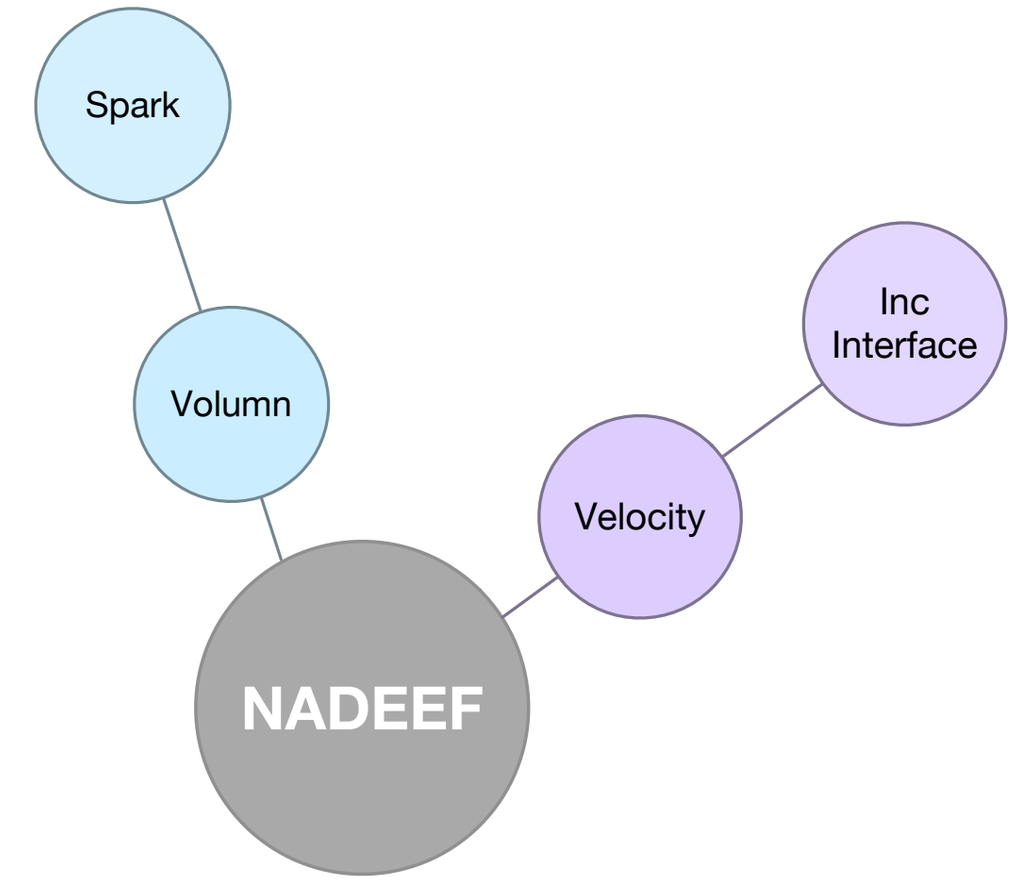


NADEEF for Big Data

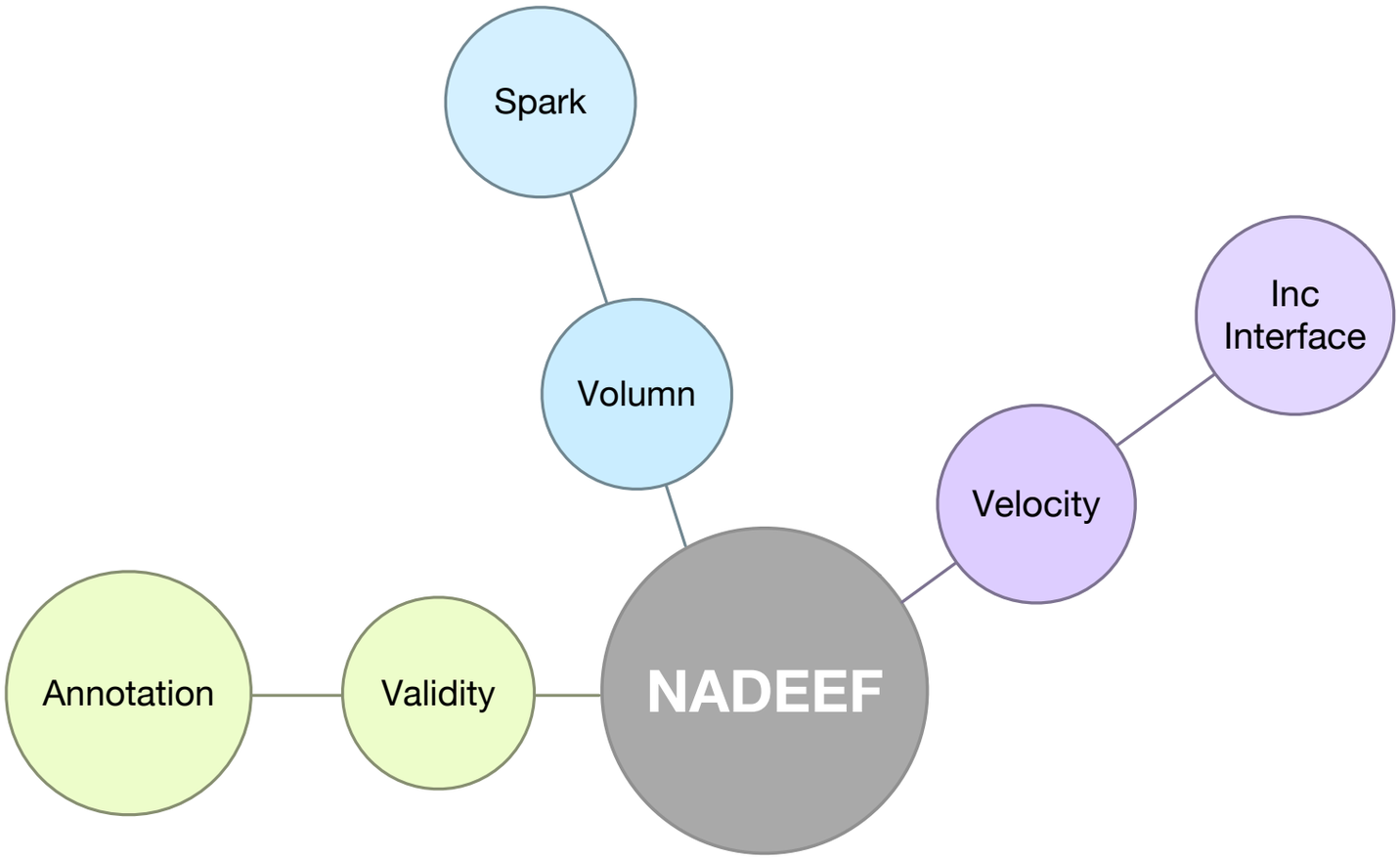
NADEEF for Big Data



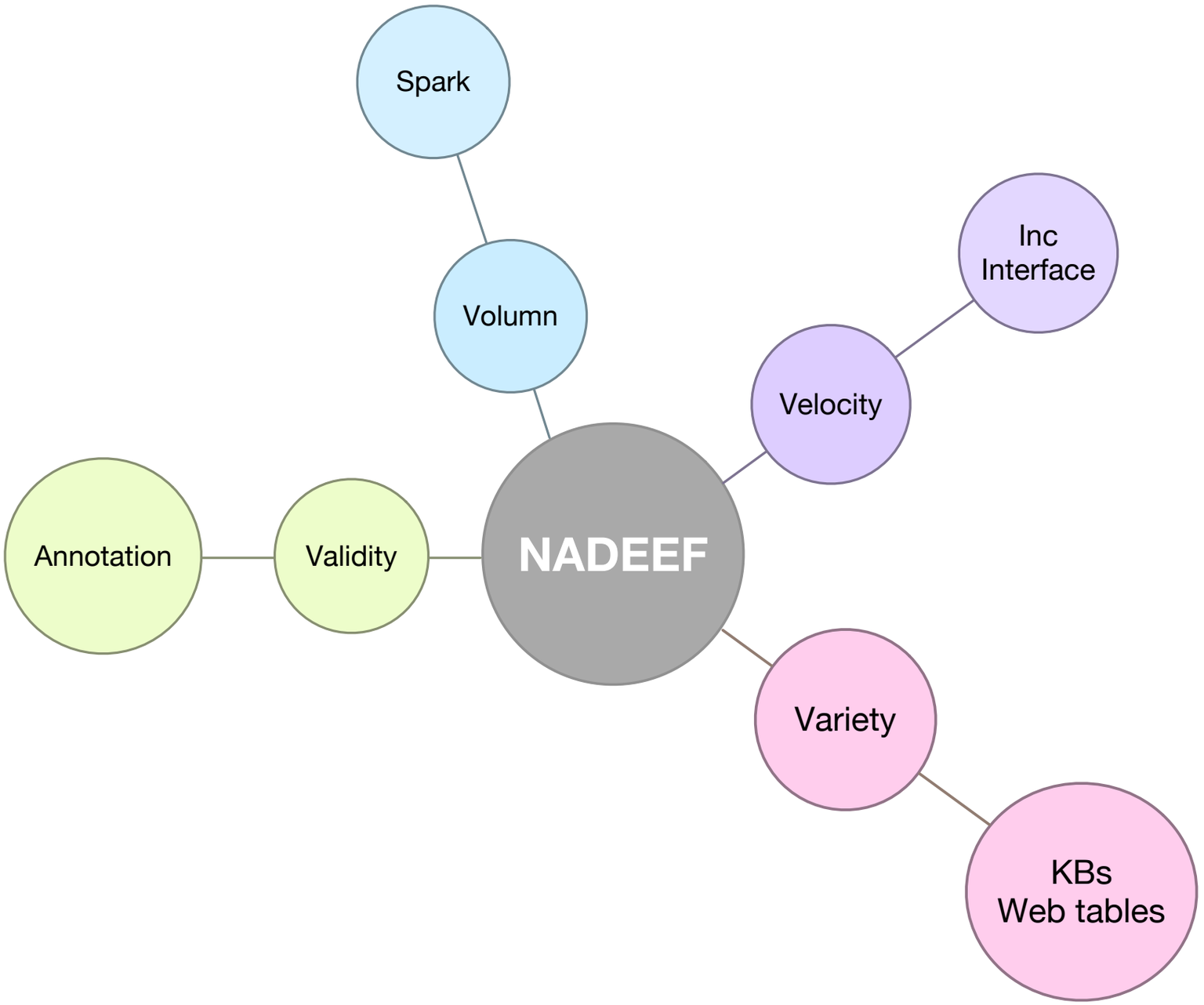
NADEEF for Big Data



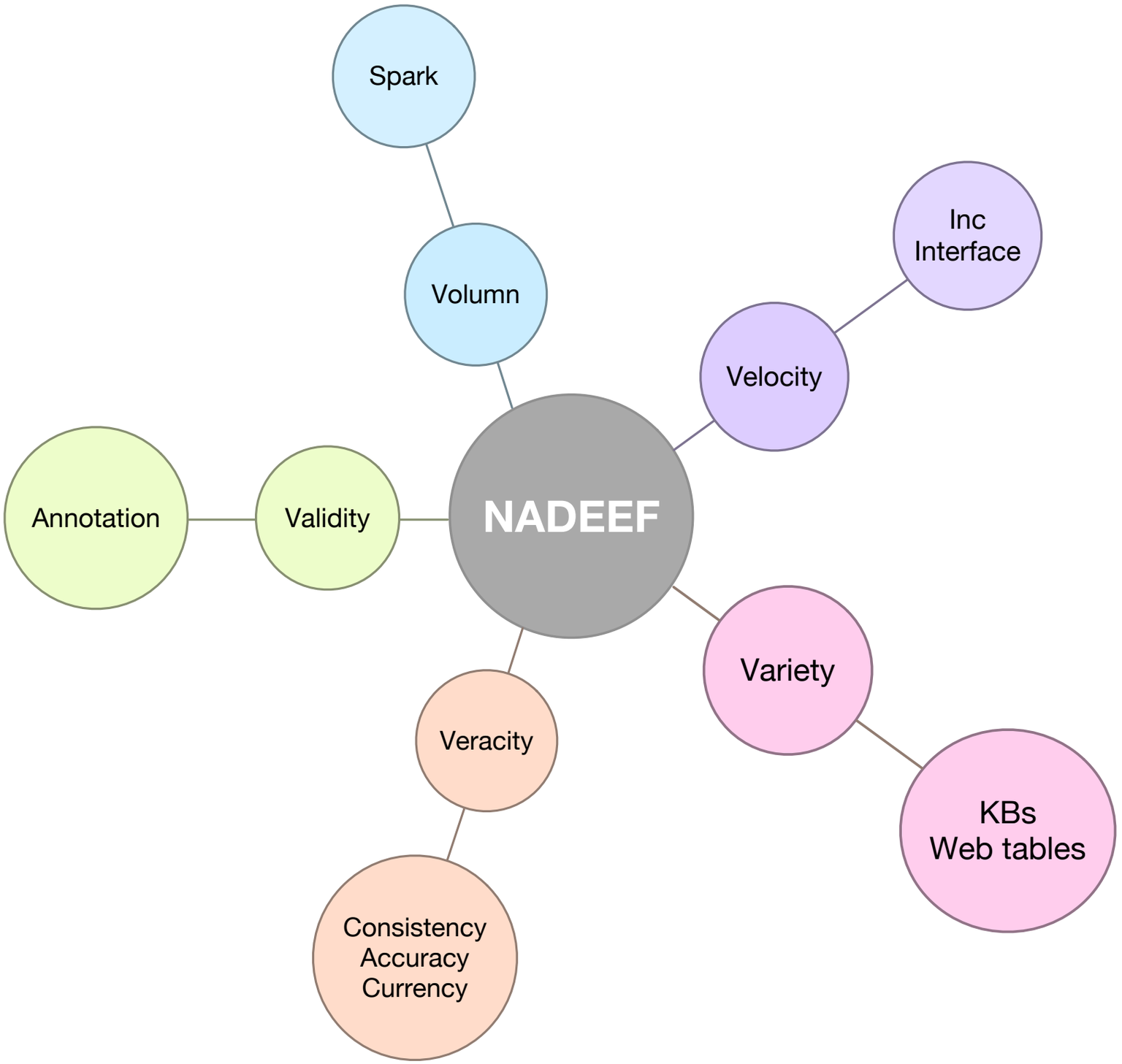
NADEEF for Big Data



NADEEF for Big Data



NADEEF for Big Data



Future Work

- Error detection
  - Rule discovery and validation
  - Combining different methods
- Explain errors to users
  - Summarization
  - Visualization
- Reliable data repairing
  - Effectively involve users as first-class citizens