CSCI-620

Data Cleaning

Data cleaning

- We can’t always rely on all of our data to be high quality
- Poor quality data affects the results of our data mining algorithms
- Data cleaning is the process of identifying “dirty” data and fixing it

Data cleaning

- What kind of data is in our dataset?
- What are the attributes and how are they related?
- Examples:
  - Nominal - labels
  - Ordinal - ordered
  - Interval - order with differences
  - Ratio - order with difference/zero

Nominal

- Names of things
- Categories
- Tags
- Genres

Ordinal

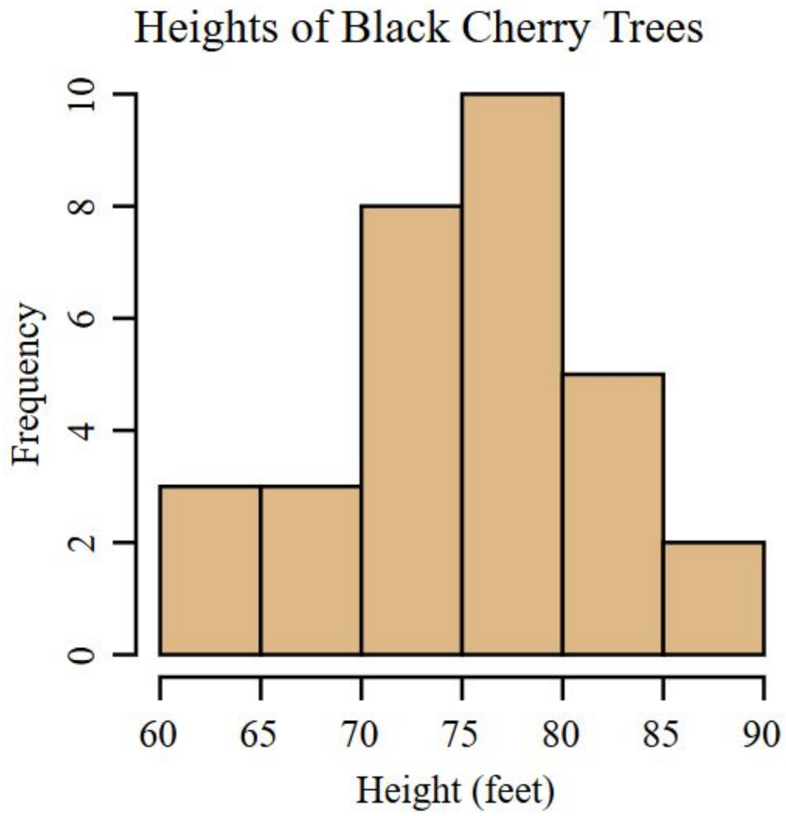
- Likert scales
- High/Medium/Low
- G/PG/PG-13/R/NC-17

Interval

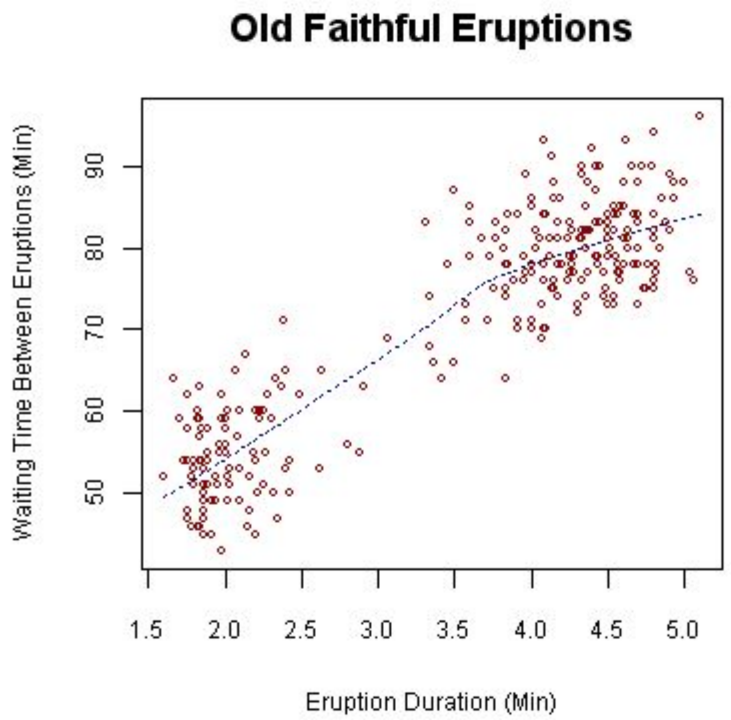
- Dates
- Times
- Temperature

Ratio

- Money
- Elapsed time
- Height/weight
- Age


Heights of Black Cherry Trees
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (feet)

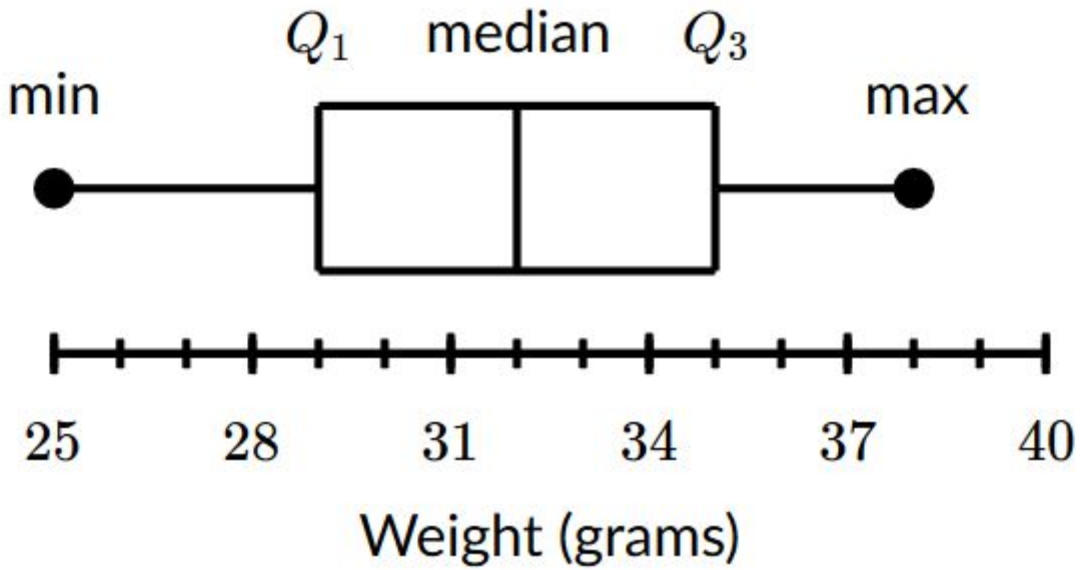
Histogram



Scatter plot


16000
13000
12000
8000
2007
2008
2009
2010
2011
2012
Source: WSJ Markets Data Group

Time series



Box and whisker

Descriptive statistics

- Central tendency
  - Where does the data centred?
- Variation
  - How spread out is the data?

Having N data points each of which takes value x_i, the mean is computed as follows:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Mean

Having N data points each of which takes value y_i, the median is computed as follows:

- N is odd:

$$\tilde{y} = y_{(N+1)/2}$$

- N is even:

$$\tilde{y} = \frac{y_{N/2} + y_{(N/2)+1}}{2}$$

Median

First quartile: middle data point between smallest and median.

Third quartile: middle data point between median and highest.

Ordered Data Set: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

	Method 1	Method 2	Method 3
Q_1	15	25.5	20.25
Q_2	40	40	40
Q_3	43	42.5	42.75

Quartiles

Having N data points each of which takes value y_i, the mode is the most frequent value.

There can be multiple modes.

Mode

Having N data points each of which takes value x_i, the variance is computed as follows:

$$s^2 = \frac{\left(x_i - \overline{x}\right)^2}{N-1}$$

Variance

$$s = \sqrt{s^2}$$

Standard deviation

Data quality

- Validity - does it meet our rules?
- Accuracy - is it correct?
- Completeness - is data missing?
- Consistency - does data match up?
- Uniformity - are we using similar units?
- Timeliness - is data up to date?

ssn	manager	salary
145-4348-71	false	120000
84-2841-91	true	90000
62-1456-31	true	130000

Managers should have higher salaries than non-managers

Validity

ssn	city	zip
145-4348-71	Boston	02134
84-2841-91	Burbank	91501
62-1456-31	Rochester	85001

Accuracy

ssn	city	zip
145-4348-71	Boston	02134
84-2841-91	Burbank	??
62-1456-31	Rochester	14623

Completeness

ssn	salary
145-4348-71	120000
84-2841-91	93000

ssn	salary
145-4348-71	150000
84-2841-91	93000

Consistency

product	volume
Coca Cola	12
Pepsi	12

fl oz

product	volume
Fanta	355
Pepsi	355

mL

Uniformity

country	leader
USA	Barack Obama
Canada	Stephen Harper

Timeliness

Other examples

- Duplicate data
- Empty rows
- Abbreviations
- Typos
- Missing values
- Extra spaces
- Incorrect types
- Stale data
- Outliers
- Uniqueness

ssn	name	salary	dept
24-37-9162	Neha Patel	120000	7
19-24-3618	Benítez, Ramón	130	1
24-37-9162	Neha Patel	120000	3
21-96-43967	Cecilia Yang	160000	

dept	name
1	Radiology
2	Emegency
3	Cardiology
4	Nephrology
5	Arkansas

Data quality examples

Quality constraints

- (Conditional) functional dependencies
- Inclusion dependencies
- Unique constraints
- Denial constraints
- Consecutive rule (e.g. check numbers)

Functional Dependency ZIP → City

$$\forall t_1, t_2 \in R: \neg(t_1.zip = t_2.zip \wedge t_1.city \neq t_2.city)$$

Denial constraints

$$\forall t_1 \in R: \neg(t_1.openingTime > t_1.closingTime)$$

Denial constraints

$$\forall t_1, t_2 \in R: \neg(t_1.state = t_2.state \wedge t_1.income > t_2.income \wedge t_1.taxRate < t_2.taxRate)$$

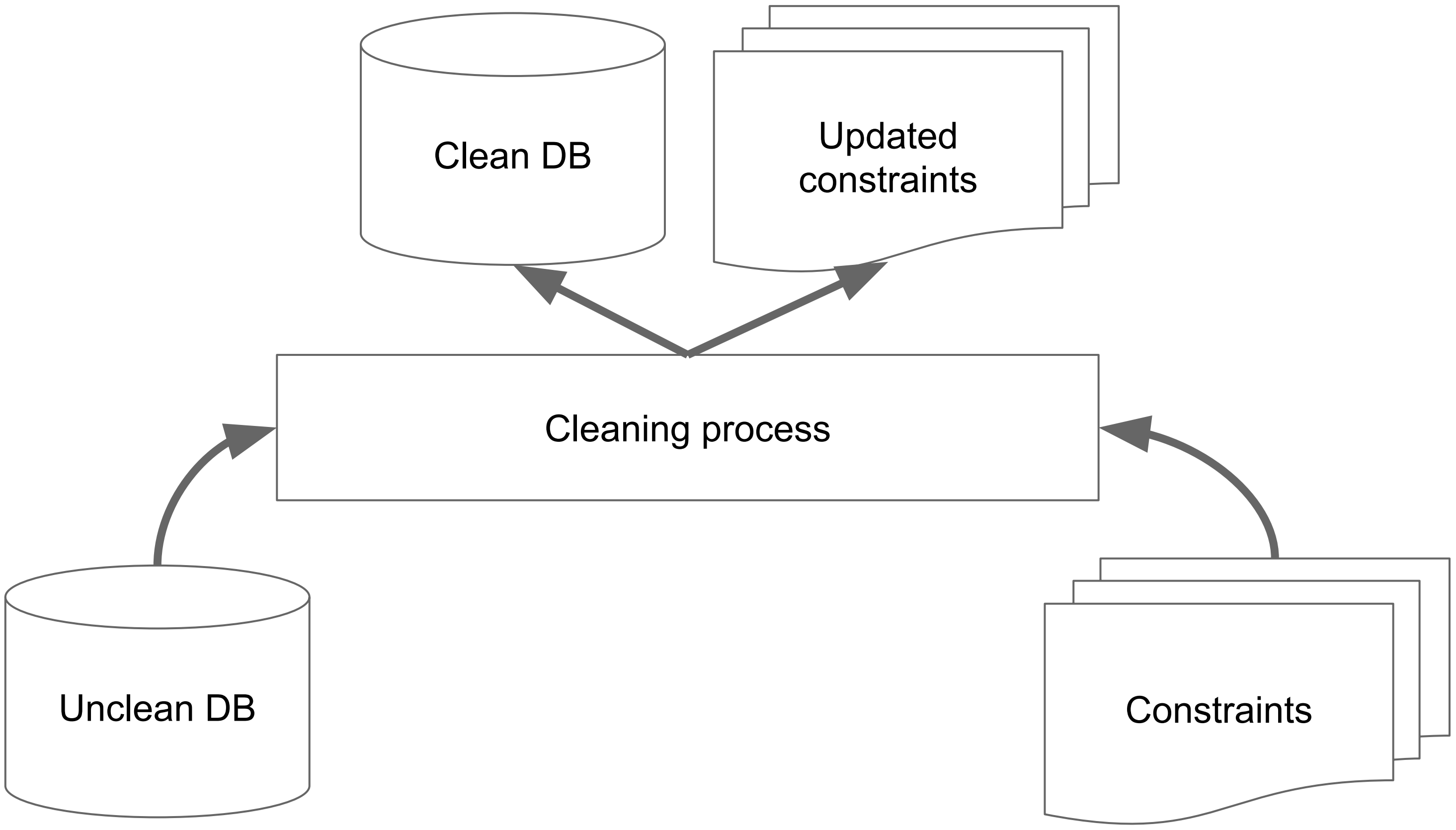
Denial constraints

LastName, FirstName → StartYear, EndYear

LastName	MiddleInitials	FirstName	StartYear	EndYear
Reagan		Ronald	1981	1989
Bush	HW	George	1989	1993
Clinton		Bill	1993	2001
Bush	W	George	2001	2009

LastName, MiddleInitials, FirstName → StartYear, EndYear

Rule evolution



Cleaning process

A	B
1	2
1	3
1	3

$A \to B$

A	B
1	3
1	3
1	3

A	B
1	2
1	2
1	2

A	B
7	2
1	3
1	3

Repairing errors

	GivenName	Surname	BirthDate	Gender	Phone	Income
t_1	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t_2	Danielle	Blake	9 Dec 1970	Female	817-988-9211	100k
t_3	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t_4	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t_5	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t_6	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

$$\text{Surname, GivenName} \rightarrow \text{Income}$$

Repairing errors

	GivenName	Surname	BirthDate	Gender	Phone	Income
t_1	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t_2	Danielle	Blake	9 Dec 1970	Female	817-988-9211	**120k**
t_3	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t_4	Hong	Li	8 Mar 1979	Female	498-214-5822	**90k**
t_5	Ning	Wu	3 Nov 1982	Male	313-134-9241	**95k**
t_6	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

$$\text{Surname, GivenName} \rightarrow \text{Income}$$

Trusted FD

	GivenName	Surname	BirthDate	Gender	Phone	Income
t_1	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t_2	Danielle	Blake	9 Dec 1970	Female	817-988-9211	100k
t_3	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t_4	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t_5	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t_6	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName, BirthDate, Phone → Income

Trusted data

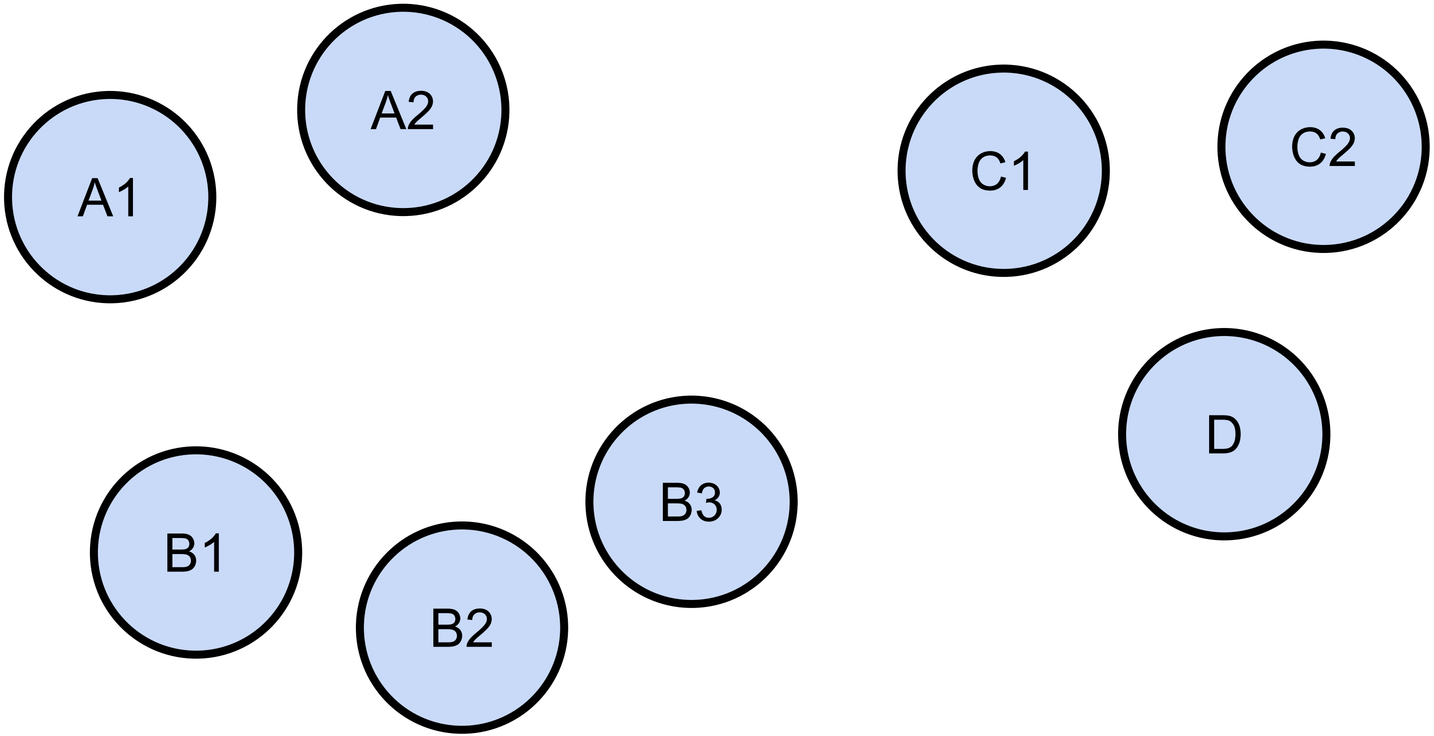
	GivenName	Surname	BirthDate	Gender	Phone	Income
t_1	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t_2	Danielle	Blake	9 Dec 1970	Female	817-988-9211	**120k**
t_3	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t_4	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t_5	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t_6	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName, BirthDate → Income

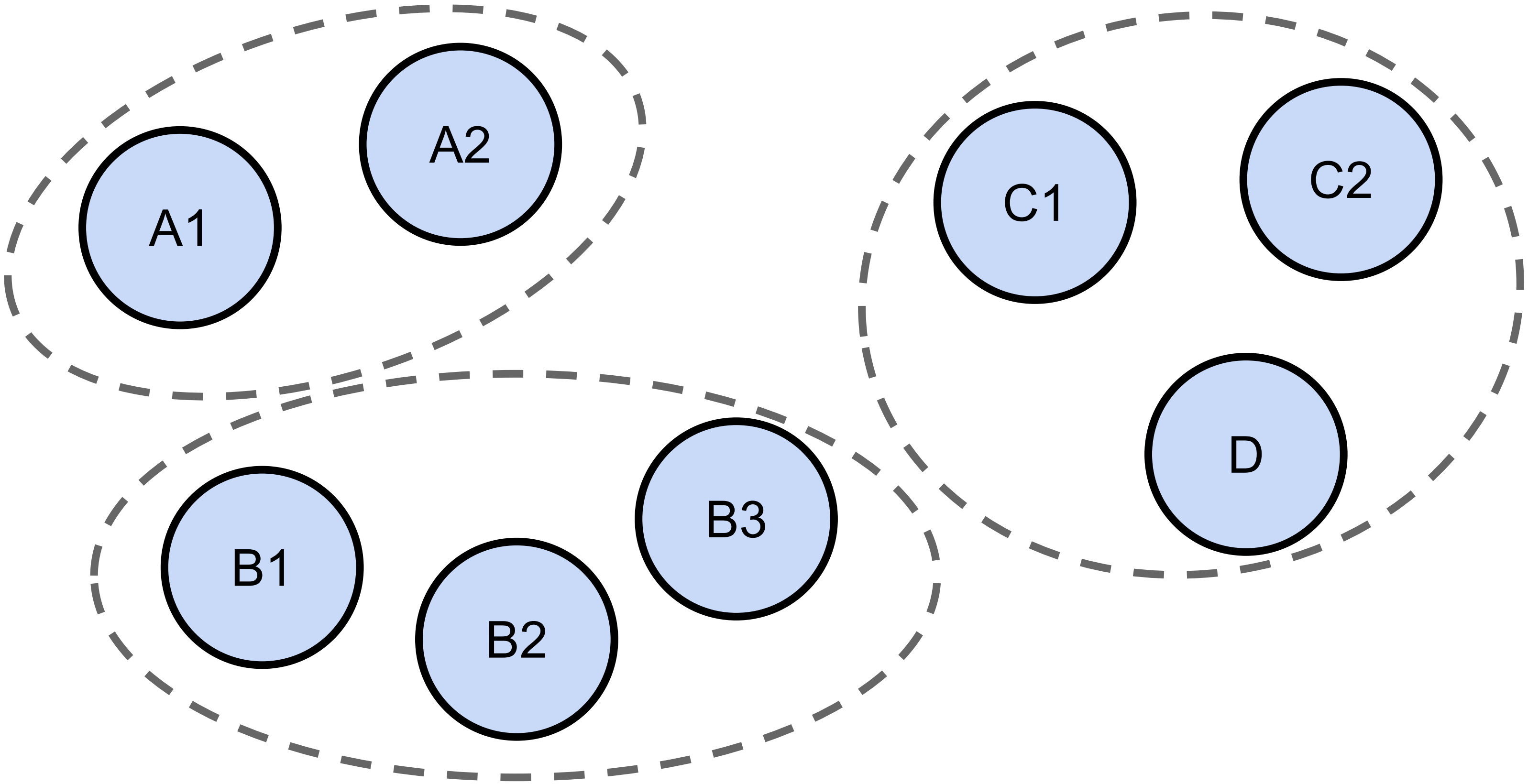
Trusted data and FD

Duplicate elimination

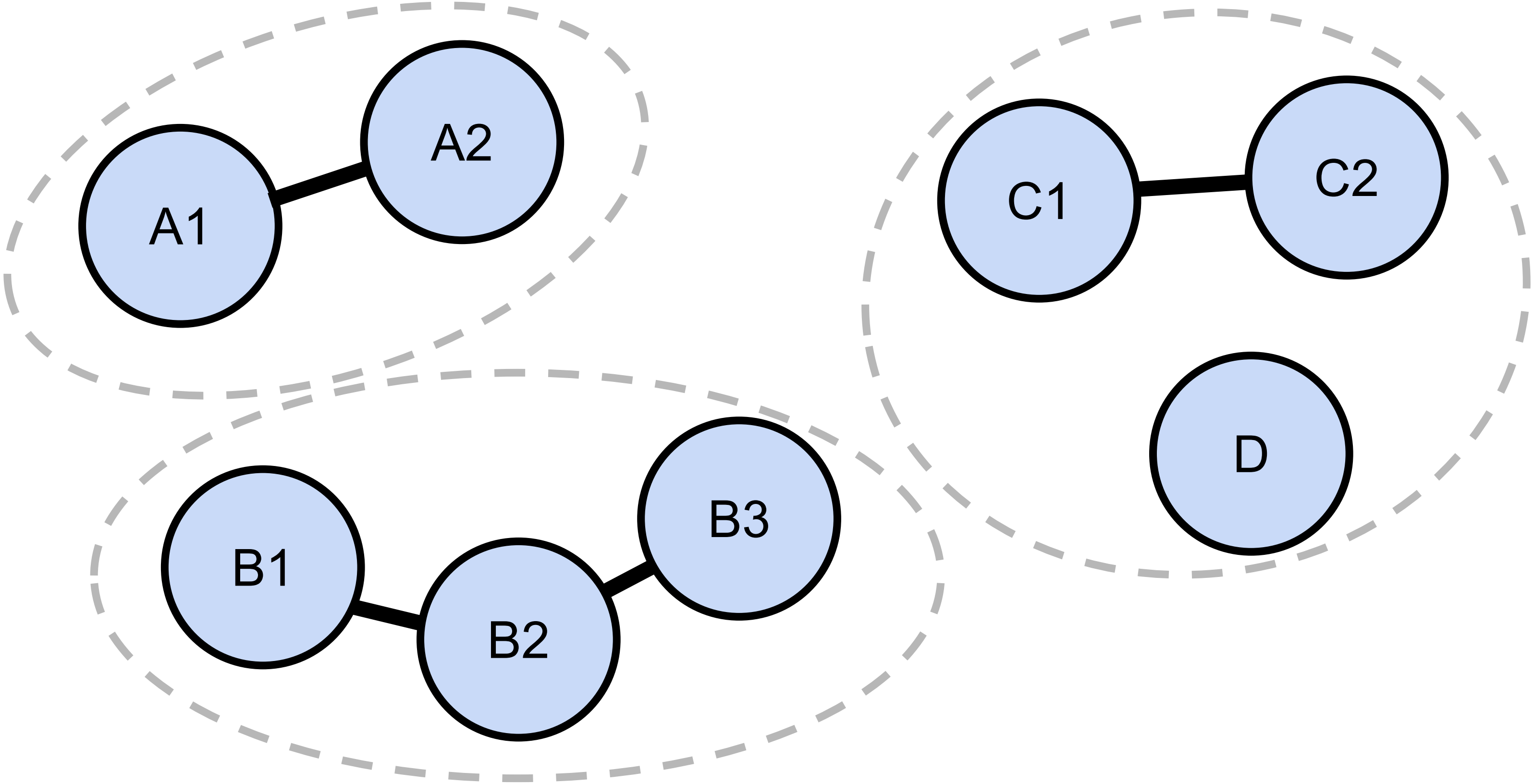
1. Blocking - find similar records
2. Pairwise matching - compare pairs
3. Clustering - combine into entities



Duplicate elimination



Duplicate elimination



Duplicate elimination

Duplicate elimination

Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k