6.S080
Data Cleaning

EXAMPLE TASK

How many people work in the US IT industry?
What is the avg revenue per employee in the tech industry?

EXAMPLE TASK

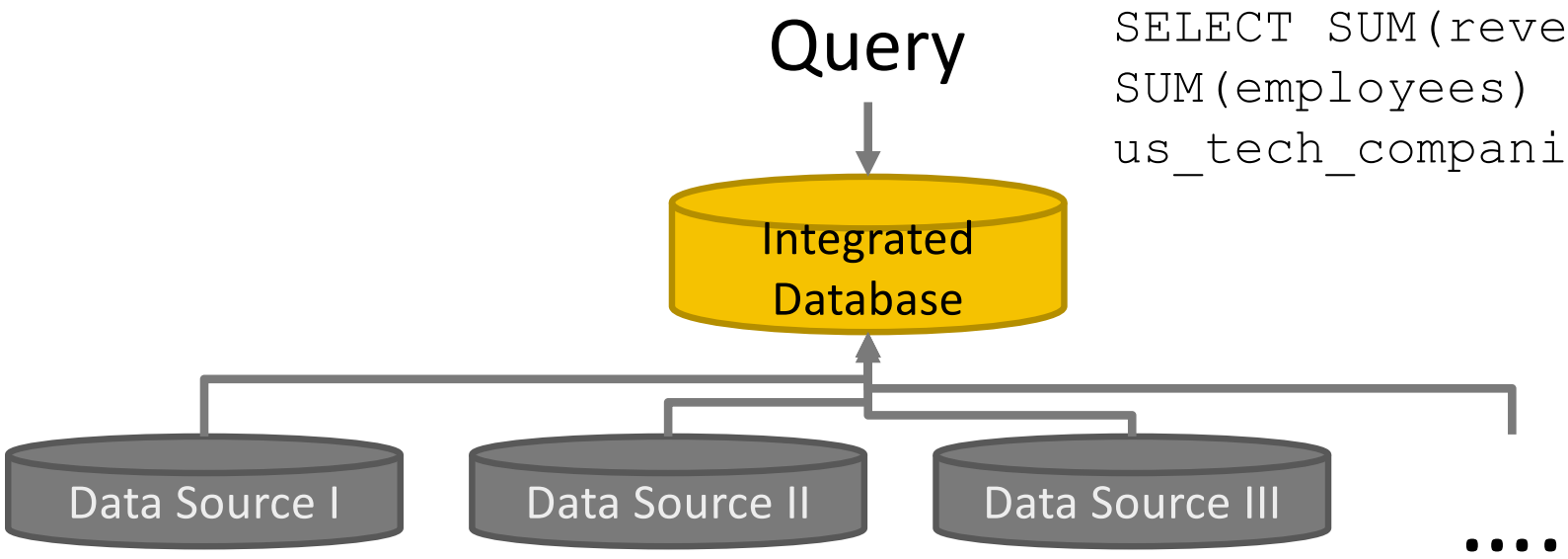
Rank[1]	Company	Fiscal Year Ending	Revenue ($B) USD	Employees	Headquarters
1	Apple Inc.	30 September 2017[2]	$229.2[1][3]	123,000[3]	Cupertino, CA, US
2	Samsung Electronics	31 December 2017[4]	$211.9[1][5][6]	320,670[7][8]	Suwon, South Korea
3	Amazon	31 December 2017[9][10]	$177.9[1][10]	613,300[11]	Seattle, WA, US
4	Foxconn	31 December 2017[12][13]	$154.7–158[1][13][14]	803,126[15]	New Taipei City, Taiwan
5	Alphabet Inc.	31 December 2017[16][17]	$110.8[1][17]	80,110[18]	Mountain View, CA, US
6	Microsoft	30 June 2017[19]	$90.0[1]	124,000[19]	Redmond, WA, US
7	Huawei	31 December 2017[20][21]	$89.3–92.5[1][21]	180,000	Shenzhen, China
8	Hitachi	31 March 2018[22]	$84.6[1]	307,275	Tokyo, Japan
9	IBM	31 December 2017[23][24]	$79.1[1]	397,800	Armonk, NY, US
10	Dell Technologies	31 January 2018[25][26]	$78.7[1][26]	145,000[25]	Round Rock, TX, US
11	Sony	31 March 2018[27]	$77.1[1][28]	117,300[27]	Tokyo, Japan
12	Panasonic	31 March 2018[29]	$72.0[1]	274,143	Osaka, Japan
13	Intel	31 December 2017[30]	$62.8[1]	102,700	Santa Clara, CA, US
14	LG Electronics	31 December 2017[31]	$54.3[1]	74,000	Seoul, South Korea
15	JD.com	31 December 2017[32]	$54.0[1]	157,831	Beijing, China
16	HP Inc.	31 October 2017[33]	$52.0[1]	49,000	Palo Alto, CA, US

Private and semipublic companies with the most employees in the world			
Rank	Employer	Country	Employees
1	Walmart	United States	2,200,000
2	China National Petroleum	China	1,382,401
3	China Post Group	China	935,191
4	State Grid	China	917,717
5	Hon Hai Precision Industry (Foxconn)	Taiwan	667,680
6	Volkswagen	Germany	664,496
7	Amazon	United States	647,500
8	Sinopec Group	China	619,151
9	Compass Group	United Kingdom	595,841
10	United States Postal Service	United States	565,802

United States Largest Private Employers (as of 2017)[1][2][3][4] [hide]			
Rank	Employer	Global number of Employees	Median annual pay
1	Walmart	2,300,000	$19,177
2	Amazon	469,800	$36,969
	Deutsche Post DHL	499,018	
3	United Parcel Service	456,415	$53,443
4	Yum! Brands	450,000	$9,111
5	Kroger	449,000	$21,075
6	Home Depot	413,000	$20,095
7	Berkshire Hathaway	377,000	$53,510 (BH directly employs c. 30 people. All the others are employed by the companies BH purchases.)
8	International Business Machines	366,000	$55,088
9	FedEx	357,000	$50,017
10	Target Corporation	345,000	$20,581
11	General Electric	313,000	$57,211
12	Walgreens Boots Alliance	290,000	$31,132
13	Starbucks	277,000	$12,754
14	Albertsons	273,000	
15	PepsiCo	263,000	$47,801
16	Wells Fargo	262,700	$60,466
17	Cognizant Technology Solutions	260,000	$31,998
18	UnitedHealth Group	260,000	$58,378
19	Lowe's	240,000	$23,905
20	AT&T	268,540	$95,814

```
,name,domain,year founded,industry,size range,locality,country,linkedin url,current employee estimate,total employee estimate
5872184,ibm,ibm.com,1911,information technology and services,10001+,"new york, new york, united states",united states,linkedin.com/company/ibm,274047,716906
4425416,tata consultancy services,tcs.com,1968,information technology and services,10001+,"bombay, maharashtra, india",india,linkedin.com/company/tata-consultancy-services,190771,341369
21074,accenture,accenture.com,1989,information technology and services,10001+,"dublin, dublin, ireland",ireland,linkedin.com/company/accenture,190689,455768
2309813,us army,goarmy.com,1800,military,10001+,"alexandria, virginia, united states",united states,linkedin.com/company/us-army,162163,445958
1558607,ey,ey.com,1989,accounting,10001+,"london, greater london, united kingdom",united kingdom,linkedin.com/company/ernstandyoung,158363,428960
3844889,hewlett-packard,hpe.com,1939,information technology and services,10001+,"palo alto, california, united states",united states,linkedin.com/company/hewlett-packard-enterprise,127952,412952
2959148,cognizant technology solutions,cognizant.com,1994,information technology and services,10001+,"teaneck, new jersey, united states",united states,linkedin.com/company/cognizant,122031,210020
5944912,walmart,walmartcareers.com,1962,retail,10001+,"withee, wisconsin, united states",united states,linkedin.com/company/walmart,120753,272827
3727010,microsoft,microsoft.com,1975,computer software,10001+,"redmond, washington, united states",united states,linkedin.com/company/microsoft,116196,276983
3300741,at&t,att.com,1876,telecommunications,10001+,"dallas, texas, united states",united states,linkedin.com/company/at&t,115188,269659
5412257,united states air force,airforce.com,1947,defense & space,10001+,"randolph, texas, united states",united states,linkedin.com/company/united-states-air-force,113997,316549
2780814,pwc,pwc.com,1998,accounting,10001+,"new york, new york, united states",united states,linkedin.com/company/pwc,111372,379447
3972223,wells fargo,wellsfargo.com,,financial services,10001+,"san francisco, california, united states",united states,linkedin.com/company/wellsfargo,109532,264101
1454663,infosys,infosys.com,1981,information technology and services,10001+,"bangalore, karnataka, india",india,linkedin.com/company/infosys,104752,215718
3221953,deloitte,deloitte.com,1900,management consulting,10001+,"new york, new york, united
```

EXAMPLE TASK




```
SELECT SUM(revenue)/
SUM(employees) FROM
us_tech_companies
```

On average what is the revenue per employee in the tech sector in the US?

CLICKER: CLICKER.CSAIL.MIT.EDU/6.S080/

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

How many different **types** of errors can you find, which could influence our result (avg revenue per employee in the US)?

a) 1-2 error types
b) 3-4 error types
c) 5-6 error types
d) 7-8
e) over 8

CLICKER: CLICKER.CSAIL.MIT.EDU/6.S080/

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

Duplicate Entities (Entity Resolution)

Pattern Violation

Outdated data / wrong data

Spelling mistakes / abbreviations

Encoding Error (nb in thousands)

Rule Violations

Missing values (known unknowns)

MORE?

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

Known Unknowns

CLICKER: CLICKER.CSAIL.MIT.EDU/6.S080/

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

Unknown Unknowns

MISSING VALUES (KNOWN UNKNOWNS)

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

WHY ARE THE VALUES MISSING?

WHY ARE THE VALUES MISSING?

- **Missing Completely at Random (MCAR)**
  - Includes missing by design. For example: Survey randomly selects questions to reduce load

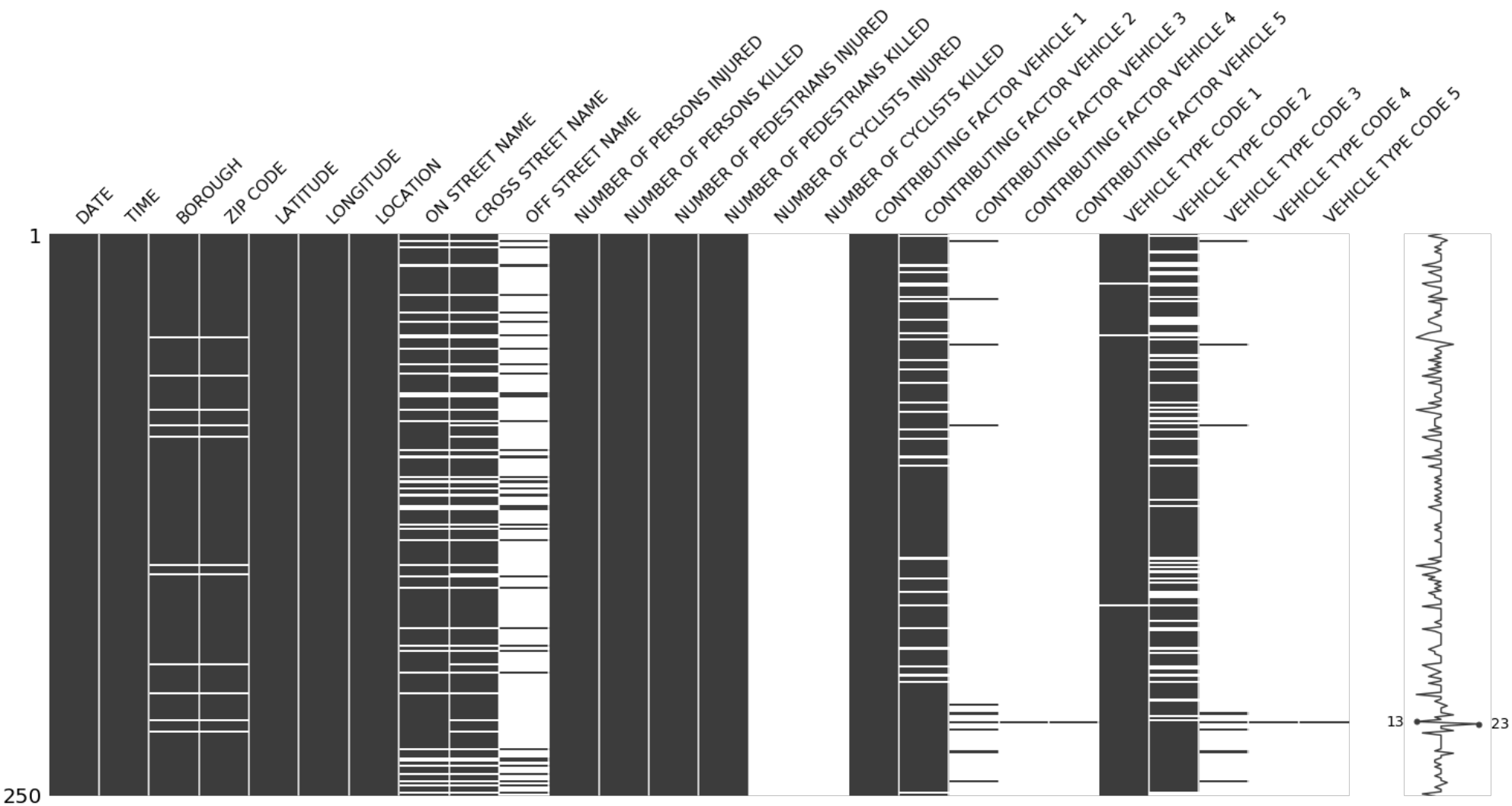
- **Missing at Random (MAR)**
  - Better name: Missing Conditionally at Random
  - Systematic relationship between the propensity of missing values and the *observed* data, but *not* the missing data.
    --> if we can control for this conditional variable, we can get a random subset.

- **Missing Not at Random, MNAR**
  - Relationship between the propensity of a value to be missing and its values.
  - Lowest education are missing on education or the sickest people are most likely to drop out of the study.
  - MNAR is called “non-ignorable” because the missing data mechanism itself has to be modeled as you deal with the missing data.

Note: null values are often encoded in various ways. Be aware of it!
Null, “null”, n/a, “”, 0, “empty”, 99999, 200.

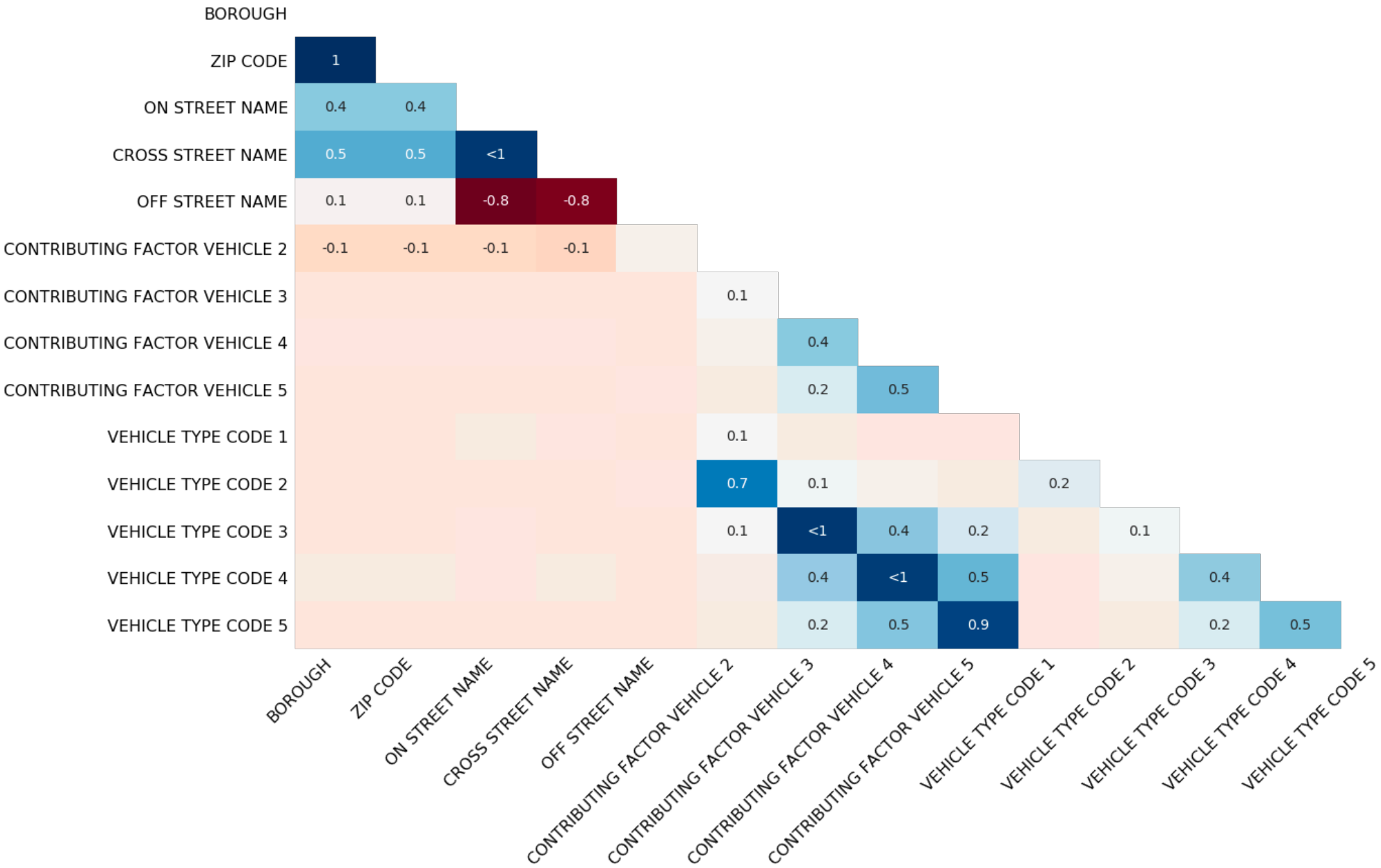
HOW DO YOU START ADDRESSING MISSING VALUES?

VISUALIZATIONS TO DETECT BIAS



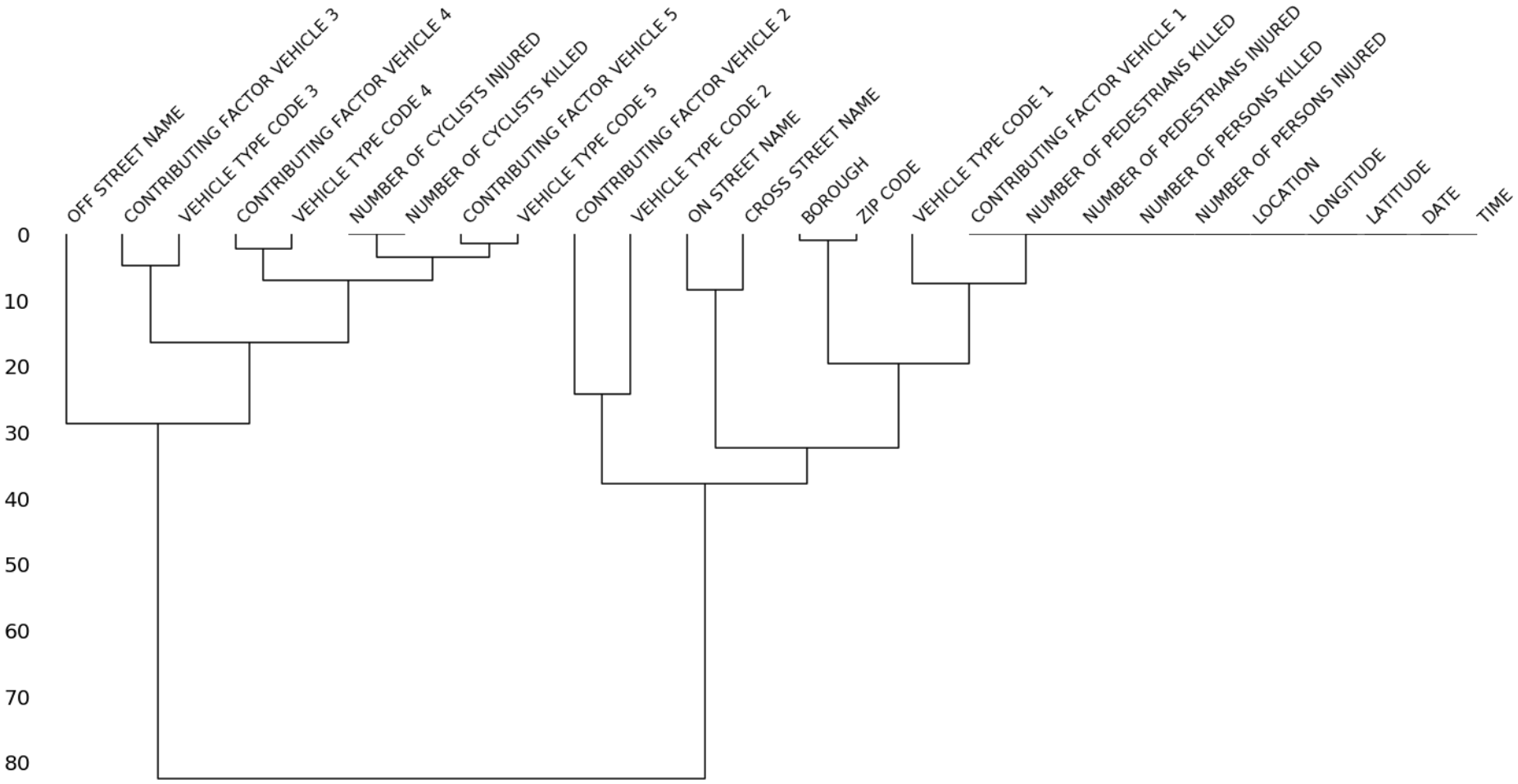
https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS



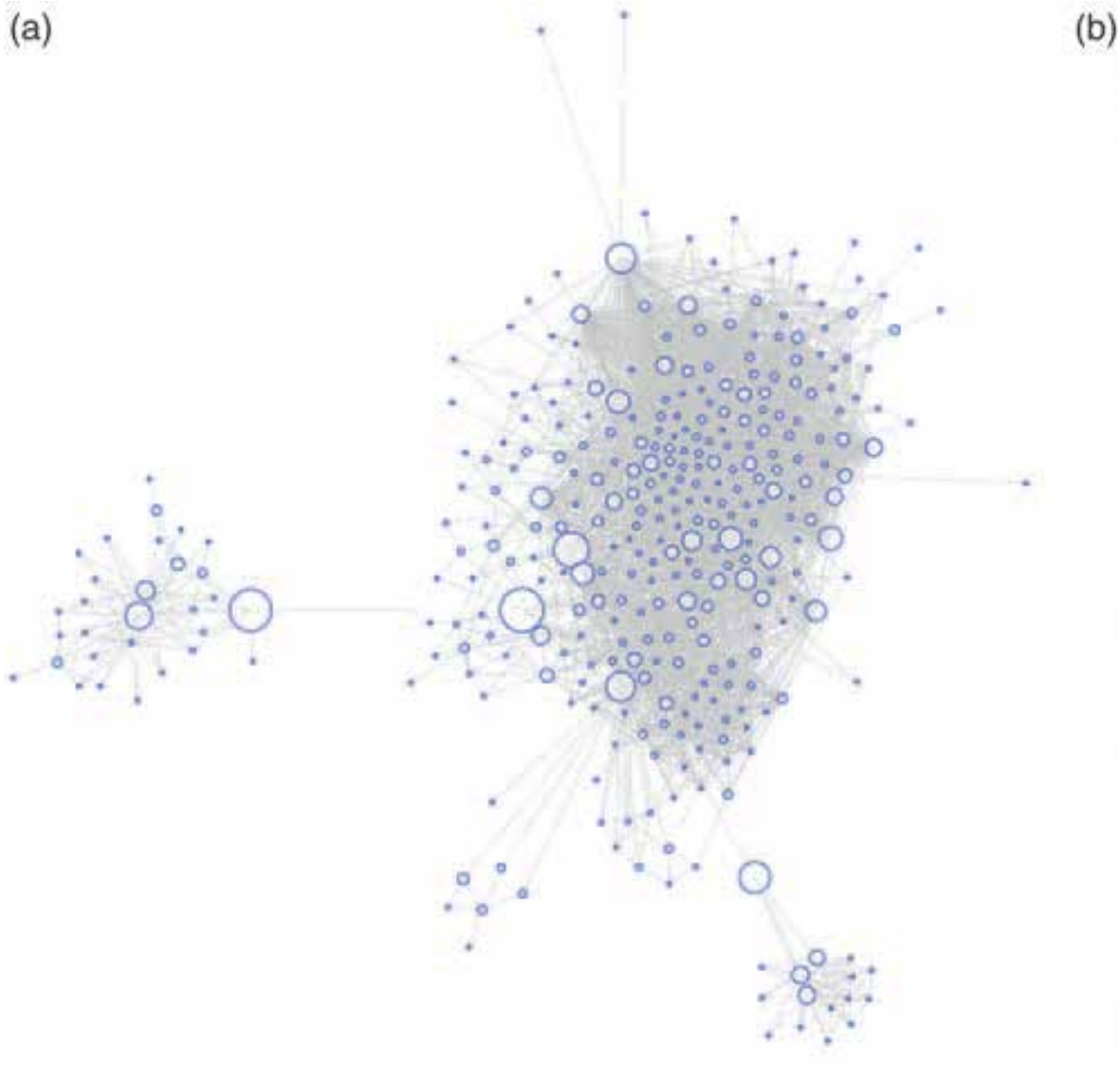
https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS



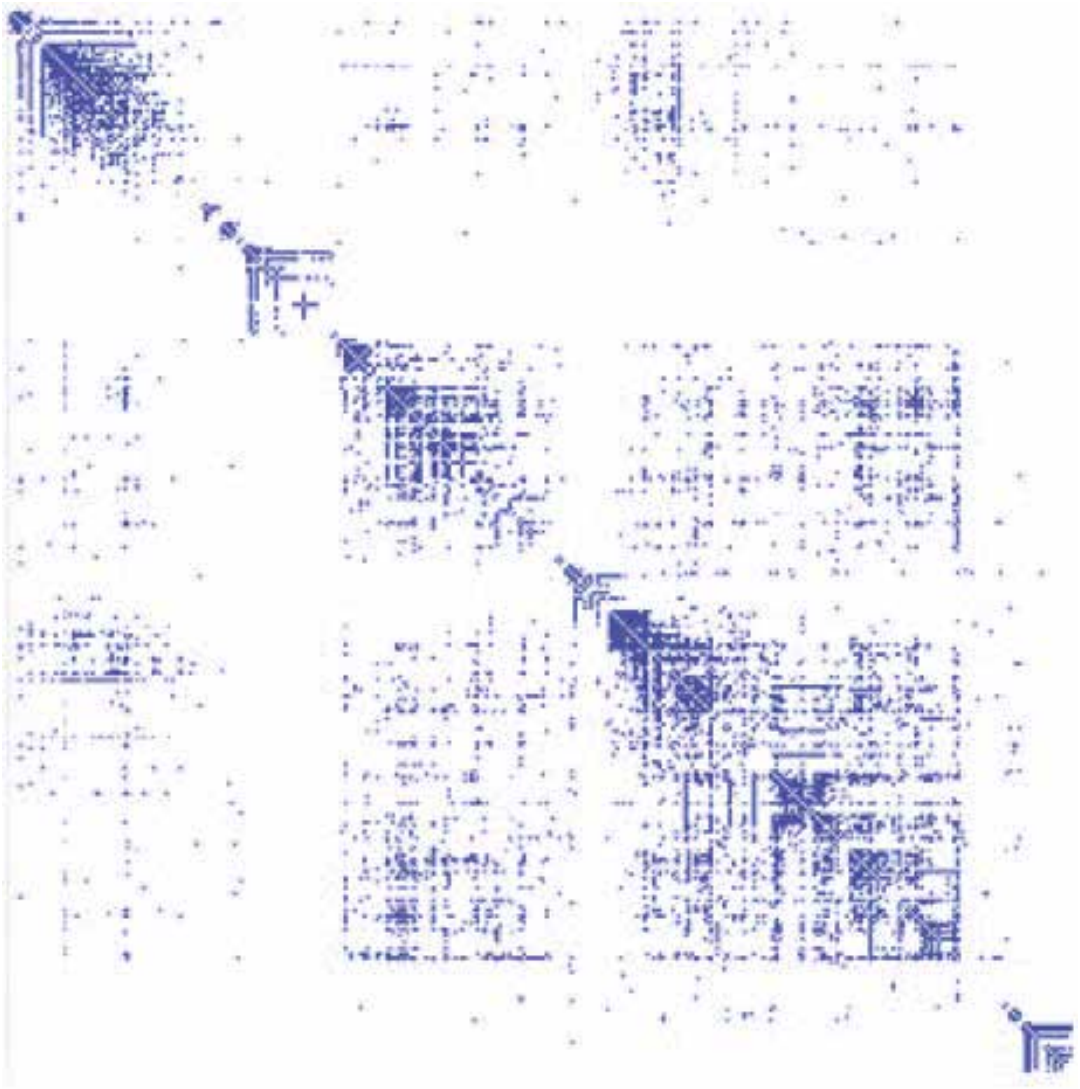
Alternative: Frequent pattern mining

FACEBOOK SOCIAL GRAPH: VISUALIZATION THE NODE-LINK DIAGRAM



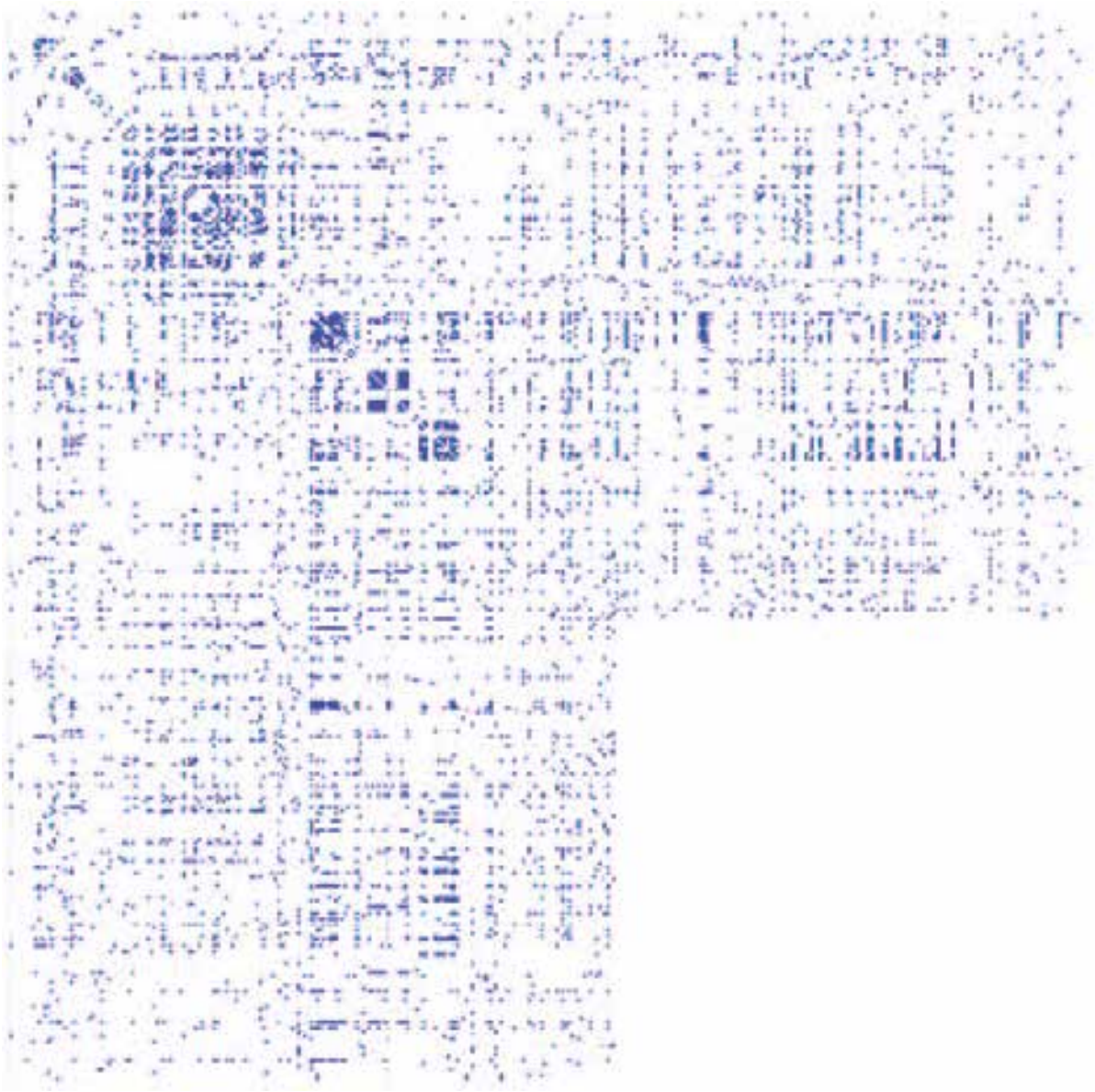
[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

FACEBOOK SOCIAL GRAPH: VISUALIZATION THE NODE-LINK DIAGRAM



[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

FACEBOOK SOCIAL GRAPH: SORTING BY RAW DATA



[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

CLASS TASK:
COME UP WITH AT LEAST 5 TECHNIQUES TO DEAL WITH MISSING VALUES

TECHNIQUES TO DEAL WITH MISSING VALUES (ONLY FOR MCAR / MAR)

- Pairwise (rarely used)
- Listwise deletion (better)
- Mean Substitution
- Dummy variable adjustment
- Maximum Likelihood Estimation
- Random sample from existing values/ reasonable distribution
- Multiple Imputation

Special cases:

- Last Observation
- Techniques for categorical values

PAIRWISE AND LISTWISE DELETION

```
SELECT SUM(revenue)/
SUM(employees) FROM
us_tech_companies
```

Pairwise Deletion

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

PAIRWISE AND LISTWISE DELETION

```
SELECT SUM(revenue)/
SUM(employees) FROM
us_tech_companies
```

Pairwise Deletion

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

Listwise Deletion

Name	Address	#Employees	Revenue	Profit
~~Google~~	~~1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA~~	~~60k~~	~~$89B~~	
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico, USA	120k	$85B	$85B
~~Tableau~~	~~Seattle, Washington, United States~~		~~$5M~~	~~$8M~~
~~Tamr~~	~~64 Church St, Cambridge, MA 02138, USA~~	~~20~~	~~$ X~~	~~$ Y~~

PAIRWISE AND LISTWISE DELETION

Pairwise Deletion

- Only cases relating to each pair of variables with missing data involved in an analysis are deleted.
- Advantage: keeps as many cases as possible for each analysis, uses all information possible with each analysis
- Disadvantage: cannot compare analyses because sample is different each time, sample size vary for each parameter estimation, can obtain nonsense results

Listwise Deletion

- Only analyze cases with available data on each variable
- Advantage: simplicity and comparability across analyses
- Disadvantage: reduces statistical power (reduced sample size), not use all information, estimates may be biased if data not MCAR

FIRST INITIAL CLEANING

Look for fields with very high percentage of missing fields

- It may be necessary to exclude field and use an alternative

Look for records with a high percentage of missing fields

- Consider excluding the case
- For example, someone who has started inputting a survey and given up after two questions!

Document that you did delete them. Very risky to forget it

UNIVARIATE SINGLE IMPUTATION
MEAN SUBSTITUTION

Mean Substitution (do not use)

- Replace missing value with the sample mean or mode. Then, run analyses as if all complete cases

UNIVARIATE SINGLE IMPUTATION
MEAN SUBSTITUTION

Mean Substitution (do not use)

- Replace missing value with the sample mean or mode. Then, run analyses as if all complete cases
- Advantage: We can use complete case analyses
- Disadvantage: Reduces variability, weakens the correlation estimates because it ignores the relationship between variables, it creates artificial band
- Unless the proportion of missing data is low, do not use this method.
- Inappropriate for categorical variables.

Dummy variable adjustment

- Create an indicator variable for missing value (1 for missing, 0 for observed), impute missing value to a constant (such as mean)

UNIVARIATE SINGLE IMPUTATION

Regression imputation

- Replace missing values with predicted score from regression equation. Use complete cases to regress the variable with incomplete data on the other complete variables.

UNIVARIATE SINGLE IMPUTATION

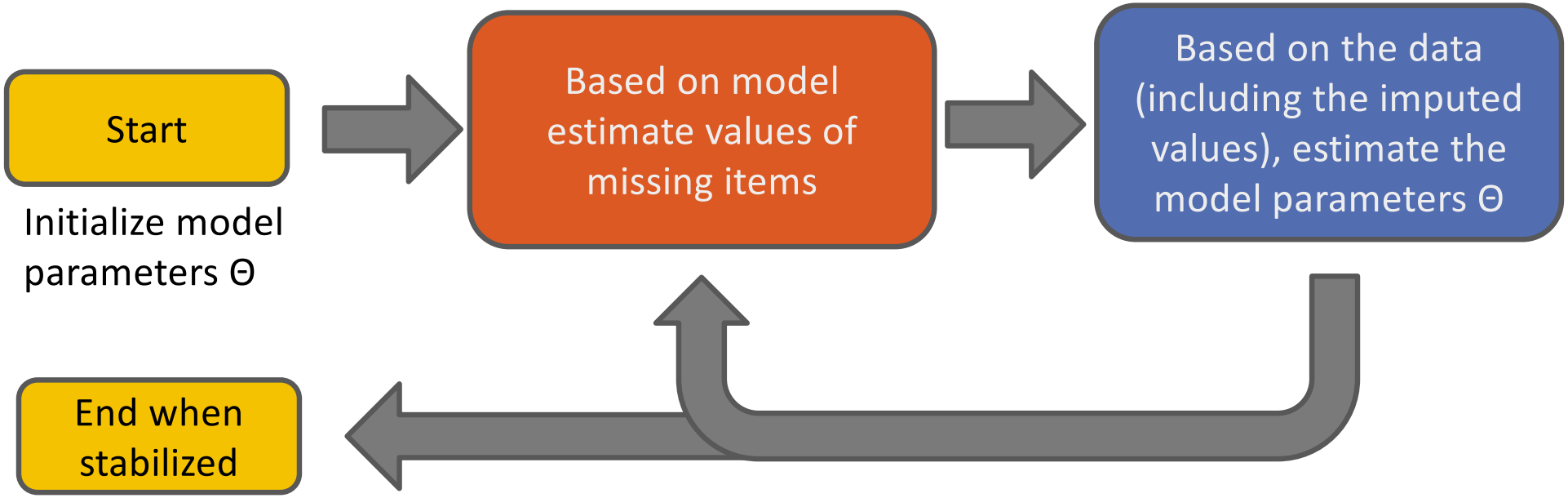
Regression imputation

- Replace missing values with predicted score from regression equation. Use complete cases to regress the variable with incomplete data on the other complete variables.
- Advantage: Uses information from the observed data, gives better results than previous ones
- Disadvantage: over-estimates model fit and correlation estimates, weakens variance

Maximum Likelihood Estimation

- Identifies the set of parameter values that produces the highest log-likelihood.

EM ALGORITHM



https://scikit-learn.org/stable/modules/impute.html

EM IMPUTATION METHODS

According to the key result of Dempster, Laird and Rubin (1977), $\theta^{(t+1)}$ is better estimate than $\theta^{(t)}$, because the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ in each iteration increases the log likelihood,

$$l\left(\theta^{(t+1)}\middle|Y_{obs}\right) \geq l\left(\theta^{(t)}\middle|Y_{obs}\right).$$

Therefore, iteration of EM algorithm can be considered in two steps: **Expectation Step** and **Maximization Step**.

E-Step: In this step, the function $Q\left(\theta\middle|\theta^{(t)}\right)$ is calculated as the conditional expectation of complete data log likelihood over the conditional predictive distribution, $f\left(Y_{mis}\middle|Y_{obs},\theta^{(t)}\right)$, of Y_{mis} given Y_{obs} and a current estimate of θ, say $\theta^{(t)}$.

M-Step: In this step, estimation of $\theta^{(t+1)}$ is carried out as if there were no missing data which is achieved by maximizing $Q\left(\theta\middle|\theta^{(t)}\right)$ from E-step.

In order to define convergency of iterations, differences of parameter estimations derived in the each iteration are considered. If the difference of consecutive estimates less than selected threshold value, then iterations are stopped. Estimations from the last iteration are used as parameter estimations.

SIMPLE STOCHASTIC IMPUTATION

Random sample from existing values:

- Randomly generate an integer from 1 to n-$n_{missing}$, then replace the missing value with the corresponding observation that you chose randomly

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States		$5M	$8M

SIMPLE STOCHASTIC IMPUTATION

Random sample from existing values:

- Randomly generate an integer from 1 to n-$n_{missing}$, then replace the missing value with the corresponding observation that you chose randomly

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	66k	$5M	$8M

- Randomly generate number between 1 and 4: Say 2 → Replace $Y_{3,5}$ by $Y_{2,3}$= 66k

SIMPLE STOCHASTIC IMPUTATION

Random sample from existing values:

- Randomly generate an integer from 1 to n-$n_{missing}$, then replace the missing value with the corresponding observation that you chose randomly

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	66k	$5M	$8M

- Randomly generate number between 1 and 4: Say 2 → Replace $Y_{3,5}$ by $Y_{2,3}$= 66k
- Disadvantage: It may change the distribution of data
- **Hot-deck approach:** draws are made from units with complete data that are 'similar' to the one with missing values (donors).

SIMPLE STOCHASTIC IMPUTATION

Random sample from existing values:

- Randomly generate an integer from 1 to n-$n_{missing}$, then replace the missing value with the corresponding observation that you chose randomly

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	$10B
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66k	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
Microsoft	Albuquerque, New Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	66k	$5M	$8M

- Randomly generate number between 1 and 4: Say 2 → Replace $Y_{3,5}$ by $Y_{2,3}$= 66k
- Disadvantage: It may change the distribution of data
- **Hot-deck approach:** draws are made from units with complete data that are 'similar' to the one with missing values (donors).

Randomly sample from a reasonable distribution

- Very similar, just based on samples from a distribution.
- For example, if gender is missing and you have the information that there re about the sample number of females and males in the population. Gender ~Ber(p=0.5) or estimate p from the observed sample
- Disadvantage: distributional assumption may not be reliable (or correct), even the assumption is correct, its representativeness is doubtful

MULTIPLE IMPUTATION (MI)

Multiple imputation (MI) one of the most attractive methods for general- purpose handling of missing data in multivariate analysis.

1. Impute missing values using an appropriate model that incorporates random variation.
2. Do this M times producing M "complete" data sets.
3. Perform the desired analysis on each data set using standard complete-data methods.
4. Average the values of the parameter estimates across the M samples to produce a single point estimate.
5. Calculate the standard errors by (a) averaging the squared standard errors of the M estimates (b) calculating the variance of the M parameter estimates across samples, and (c) combining the two quantities using a simple formula

LAST OBSERVATION CARRIED FORWARD

- This method is specific to time or longitudinal data problems.
- For each individual, NAs are replaced by the last observed value of that variable. Then, analyze data as if data were fully observed.
- Disadvantage: The covariance structure and distribution change seriously

Cases	1	2	3	4	5	6
1	3.8	3.1	2.0	2.0	2.0	2.0
2	4.1	3.5	2.8	2.4	2.8	3.0
3	2.7	2.4	2.9	3.5	3.5	3.5

CATEGORICAL VALUES

Extra category

- This is bad practice because in many statistical analysis
- the impact of this strategy depends on how missing values are divided among the real categories, and how the probability of a value being missing depends on other variables;
- very dissimilar classes can be lumped into one group;
- severe bias can arise, in any direction, and when used to stratify for adjustment (or correct for confounding) the completed categorical variable will not do its job properly.

Better techniques:

- Maximum Likelihood Estimation
- KNN
- Stochastic variants

CLICKER

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA	50000	$100000M	$40000M
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	70000	$200000M	$50000M
IBM	1 New Orchard Rd; 10504	400000	$100000M	null
Microsoft	Albuquerque, New Mexico	130000	$125000M	$40000M
Tableau	Seattle, Washington, United States	4000	$1000M	null
Tamr	64 Church St, Cambridge, MA 02138, USA	30	$10M	$1M
Einblick Analytics	null	8	$0.01M	$0M
Determined AI	California	15	null	$0.01M

Calculate the result for `SELECT SUM(revenue)/SUM(employees) FROM s_tech_companies`
With listwise deletion, mean and linear regression substitution

For this example, which technique to deal with null values leads to the lowest revenue per employee value:

a) Listwise deletion
b) Mean substitution
c) Regression imputation

CLICKER

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA	50000	$100000M	$40000M
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	70000	$200000M	$50000M
~~IBM~~	~~1 New Orchard Rd; 10504~~	~~400000~~	~~$100000M~~	~~null~~
Microsoft	Albuquerque, New Mexico	130000	$125000M	$40000M
~~Tableau~~	~~Seattle, Washington, United States~~	~~4000~~	~~$1000M~~	~~null~~
Tamr	64 Church St, Cambridge, MA 02138, USA	30	$10M	$1M
~~Einblick Analytics~~	~~null~~	~~8~~	~~$0.01M~~	~~$0M~~
~~Determined AI~~	~~California~~	~~15~~	~~null~~	~~$0.01M~~

Calculate the result for `SELECT SUM(revenue)/SUM(employees) FROM s_tech_companies` with

a) Listwise deletion: $425B / $250k = $1.7M per employee
b) Mean substitution:
c) Regression imputation

CLICKER

Name	Address	#Employees	Revenue (M)	Profit (M)
Google	1600 Amphitheatre Parkway, Mountain View, CA	50000	$100000M	$40000M
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	70000	$200000M	$50000M
IBM	1 New Orchard Rd; 10504	400000	$100000M	null
Microsoft	Albuquerque, New Mexico	130000	$125000M	$40000M
Tableau	Seattle, Washington, United States	4000	$1000M	null
Tamr	64 Church St, Cambridge, MA 02138, USA	30	$10M	$1M
Einblick Analytics	null	8	$0.01M	$0M
Determined AI	California	15	$75000M	$0.01M

Calculate the result for `SELECT SUM(revenue)/SUM(employees) FROM s_tech_companies` with

a) Listwise deletion: $425B / $250k = $1.7M per employee
b) Mean substitution: $600B / 654k = $0.92M per employee
c) Regression imputation

CLICKER

Name	Address	#Employees	Revenue (M)	Profit (M)
Google	1600 Amphitheatre Parkway, Mountain View, CA	50000	$100000M	$40000M
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	70000	$200000M	$50000M
IBM	1 New Orchard Rd; 10504	400000	$100000M	null
Microsoft	Albuquerque, New Mexico	130000	$125000M	$40000M
Tableau	Seattle, Washington, United States	4000	$1000M	null
Tamr	64 Church St, Cambridge, MA 02138, USA	30	$10M	$1M
Einblick Analytics	null	8	$0.01M	$0M
Determined AI	California	15	$55000M	$0.01M

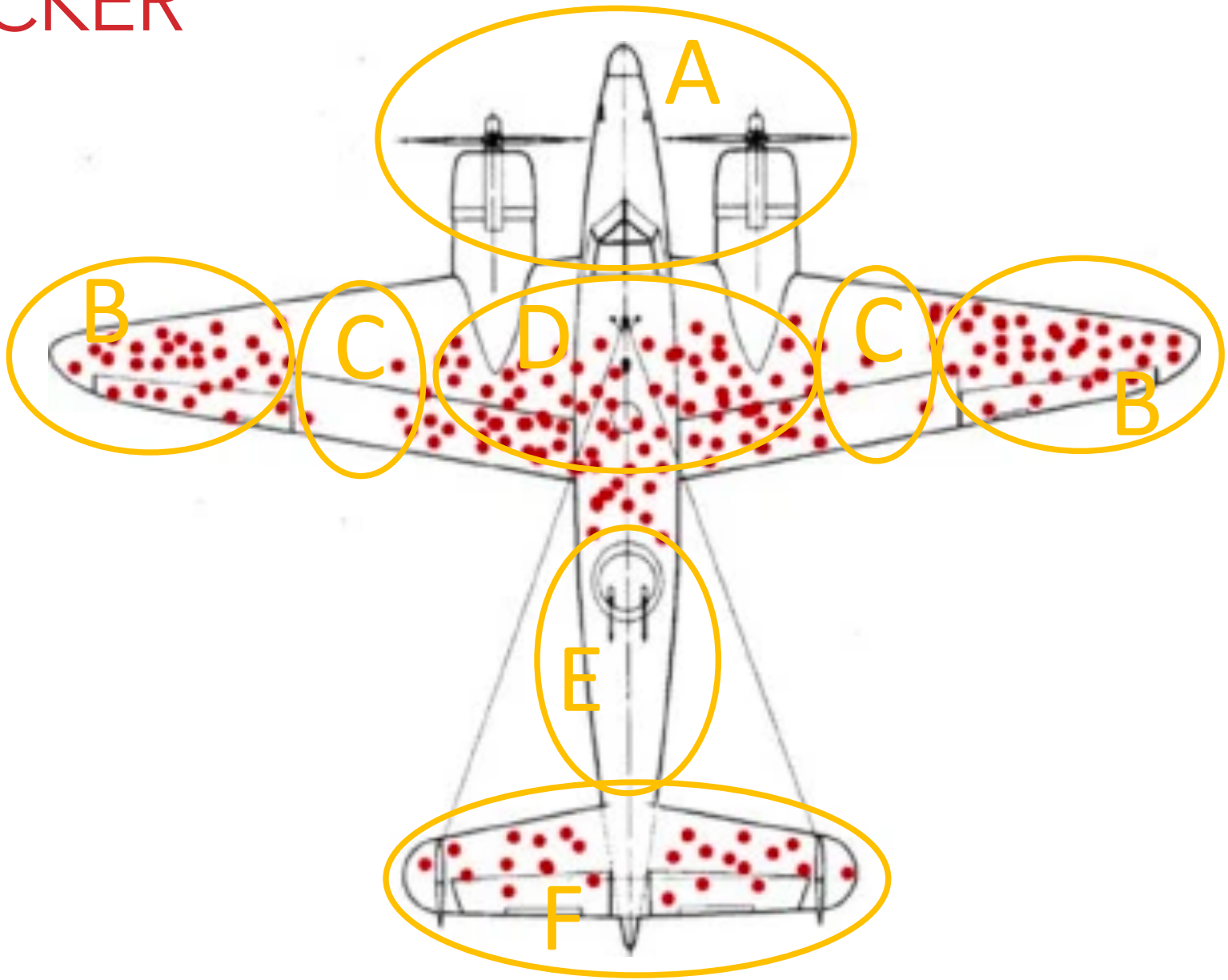
Calculate the result for `SELECT SUM(revenue)/SUM(employees) FROM s_tech_companies` with

a) Listwise deletion: $425B / $250k = $1.7M per employee
b) Mean substitution: $600B / 654k = $0.92M per employee
c) Regression imputation: $580B /654k = $0.89M per employee

Rev = 55346 + 0.212 * emp

|

CLICKER

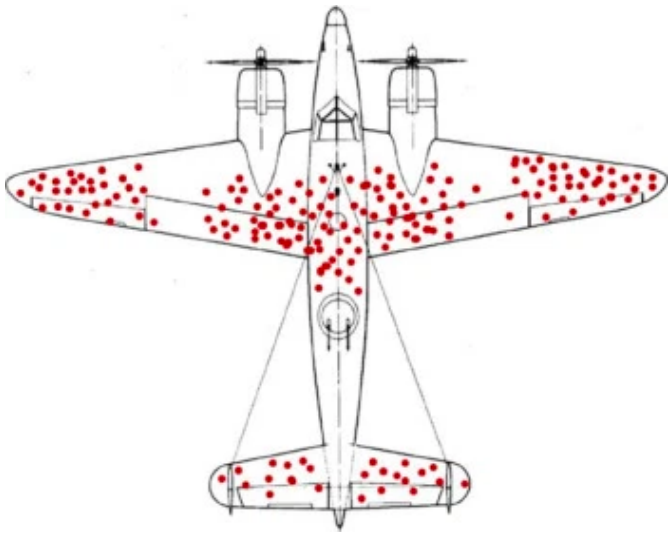


Where would you enforce the plane?

UNKNOWN UNKOWNS

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	1 New Orchard Rd; 10504	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

IF YOU CAN ESTIMATE THEM DEPENDS ON THE SAMPLING SCENARIO



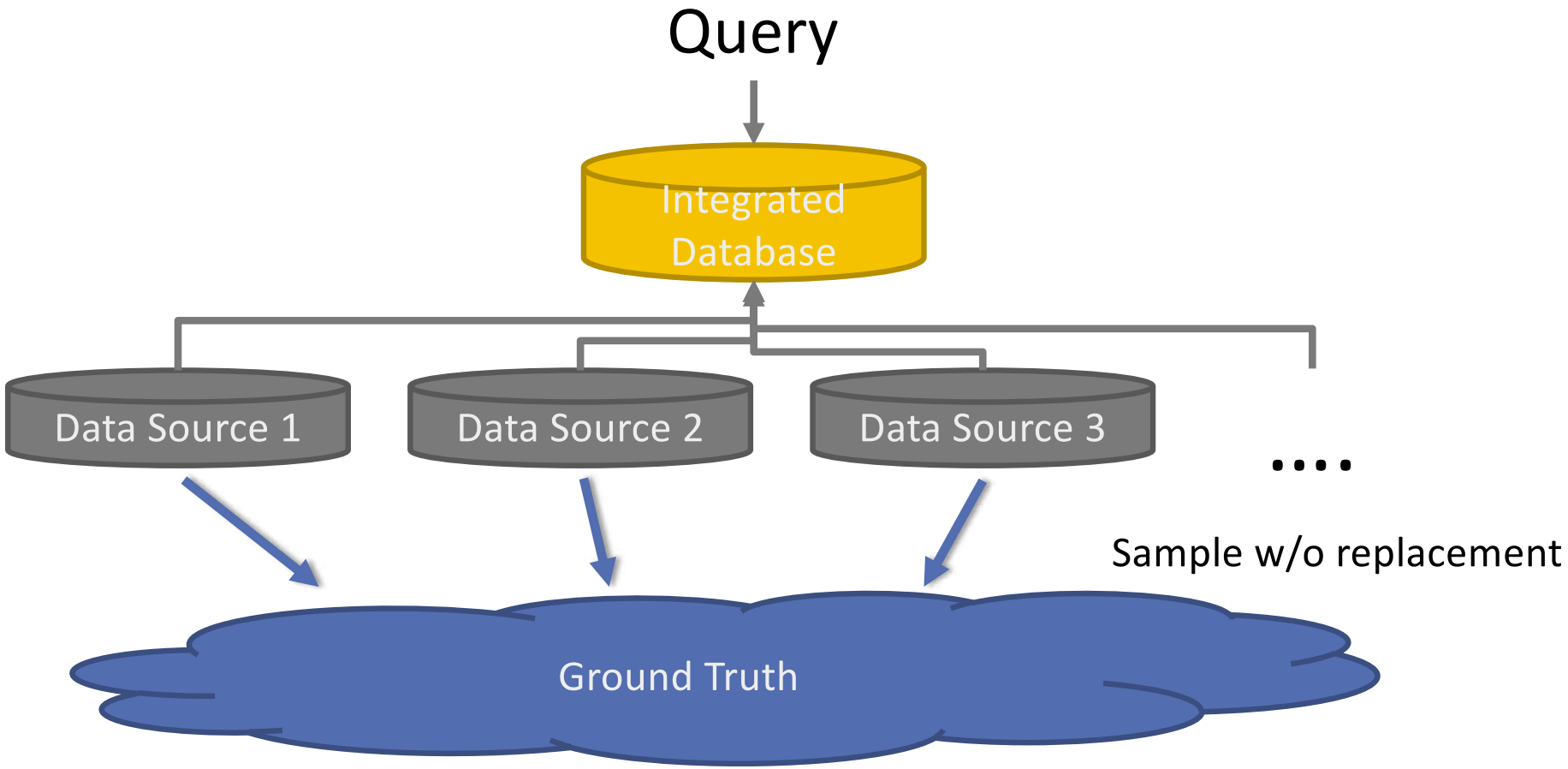
VS

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	1 New Orchard Rd; 10504	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y
Amazon	??	??	??	??
Facebook	??	??	??	??
??	??	??	??	??
??	??	??	??	??

THE IMPACT OF THE **UNKNOWN UNKNOWNS** ON QUERY RESULTS

How many people work in the US IT industry

```
SELECT SUM(employees)
FROM us_tech_companies
```



Assumption: Enough data sources , Data sources are (semi-) independent

Sampling - Statistic

$\sum$

Name	Address	#Employees	Revenue	Profit	Frequency
Google	Address I	60k	$89B	$10B	5
Apple	Address II	66k	$215B	$45B	4
IBM	Address II	380k	$80B	$12B	4
Microsoft	Address	120k	$85B	$85B	5
Tableau	Address	3.2k	$500	$8M	2
Tamr	Address	20	$-X	$-Y	1

Fingerprint (i.e., f-statistic):

f_1: 1 tamr

f_2: 1 tableau

f_4: 2 IBM

f_5: 2 Google Microsoft

Singletons (items which were exactly observed once)

MANY WAYS TO ESTIMATE THE NUMBER OF MISSING ITEMS

- Good-Turing Estimate / Chao84
- Chao92
- Pattern Maximum Likelihood
- Linear programming-based solutions (see Valiant brothers)
- …

ESTIMATING THE NUMBER OF DISTINCT BUTTERFLY SPECIES

17500 **species** known in the world

GOOD-TURING / CHAO84 ESTIMATE

Unique Items

$$\widehat{N} = \frac{c}{\left(1 - {f_1}/{n}\right)}$$

Missing mass

Number of Unknown Unknowns:

$$M = \widehat{N} - c$$

*Note, we usually prefer **Chao92**: A. Chao and S. Lee, "**Estimating the Number of Classes via Sample Coverage**," Journal of the American Statistical Association, vol. 87, no. 417, pp. 210–217, 1992*
*over **Chao84**: A. Chao, "**Nonparametric Estimation of the Number of Classes in a Population**," SJS, vol. 11, no. 4, 1984*

A NAÏVE ESTIMATOR FOR THE IMPACT OF THE UNKNOWN UNKNOWNS

```
SELECT SUM(employees)
FROM us_tech_companies
```

$$\sum employees\,,\ \Delta(employees, fingerprint)$$

$$\Delta_{Naive} = M \cdot \emptyset$$

Estimate of Unknown Unknowns Count

Average Value of Knowns (aka mean substitution)

A NAÏVE ESTIMATOR FOR THE IMPACT OF THE UNKNOWN UNKNOWNS

Number of unique records i.e., count(*)

Value sum over all unique items

$$\Delta_{Naive} = \frac{c}{\left(1 - {}^{f_1}/_{n}\right)} \cdot \frac{\sum_{\{c\}} v}{c}$$

Estimated number of missing records

Mean value

EXAMPLE

MIT Fan DB

FanID	Name	Address	Email	FanOf	Genre
2	Tim	46 Pumpkin St	timk	Nickelback, Creed, Limp Bizkit	Terrible
3	Matt	Vassar Str	Mattp	Nickelback	Terrible

MIT CSAIL DB

ID	Name
10	Tim
14	Matt

MIT Department DB

ID	Name
10	Tim
14	Joana

FanID	Name	Address	Email	FanOf	Genre	Frequency
2	Tim	46 Pumpkin St	timk	Nickelback, Creed, Limp Bizkit	Terrible	3
3	Matt	Vassar Str	Mattp	Nickelback	Terrible	2
4	Joana					1

$$\#\text{Missing} = \frac{c}{\left(1 - {}^{f_1}/_{n}\right)} = \frac{3}{(1 - {}^{1}/_{6})} = 3.6$$

Note estimator shouldn't be used if sample coverage is below 80% (1 – f1/n) and such a small number of data sources (independent samples)

EXAMPLE

$$\#\text{Missing} = \frac{c}{\left(1 - {}^{f}1/_{n}\right)} = \frac{3}{\left(1 - {}^{1}/_{6}\right)} = 3.6$$

FanID	Name	Address	Email	FanOf	Genre	Frequency
2	Tim	46 Pumpkin St	timk	Nickelback, Creed, Limp Bizkit	Terrible	3
3	Matt	Vassar Str	Mattp	Nickelback	Terrible	2
4	Joana			Cold Play	OK	1

EXAMPLE

$$\#\text{Missing} = \frac{c}{\left(1-{}^{f}1/_{n}\right)} = \frac{3}{(1-{}^{1}/_{6})} = 3.6$$

FanID	Name	Address	Email	FanOf	Genre	Frequency
2	Tim	46 Pumpkin St	timk	Nickelback, Creed, Limp Bizkit	Terrible	3
3	Matt	Vassar Str	Mattp	Nickelback	Terrible	2
4	Joana			Cold Play	OK	1
….	…	..	…	…	…	…
5	Sam	Christmas St	Samm	Celine Dion	As cheesy as deep-fried camembert [1]	

[1] https://www.telegraph.co.uk/music/concerts/cheesy-deep-fried-camembert-celine-dion-o2-arena-review/

WRONG DATA: RULE-BASED APPROACHES

Name	Address	#Employees	Revenue	Profit
Google	1600 Amphitheatre Parkway, Mountain View, CA, 94043, USA	60k	$89B	null
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	66	$215B	$45B
IBM	1 New Orchard Rd; New York 10504, USA	380k	$80B	$12B
International Business Machine	10504; 1 New Orchard Rd	380k	$-999B	$12B
Microsoft	Albuquerque, Mexico	120k	$85B	$85B
Tableau	Seattle, Washington, United States	-	$0.9B	$1B
Tamr	64 Church St, Cmabridge, MA 02138, United States	20	null	$-Y

Outdated data / wrong data

Spelling mistakes / abbreviations

Encoding Error
(nb in thousands)

Rule Violations

TWO COMPONENTS

1. Detection
2. Repair
   - Detection techniques can be used for repair
   - Missing value techniques

ERROR DETECTION

FD: [country] -> [capital]

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

emp

	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

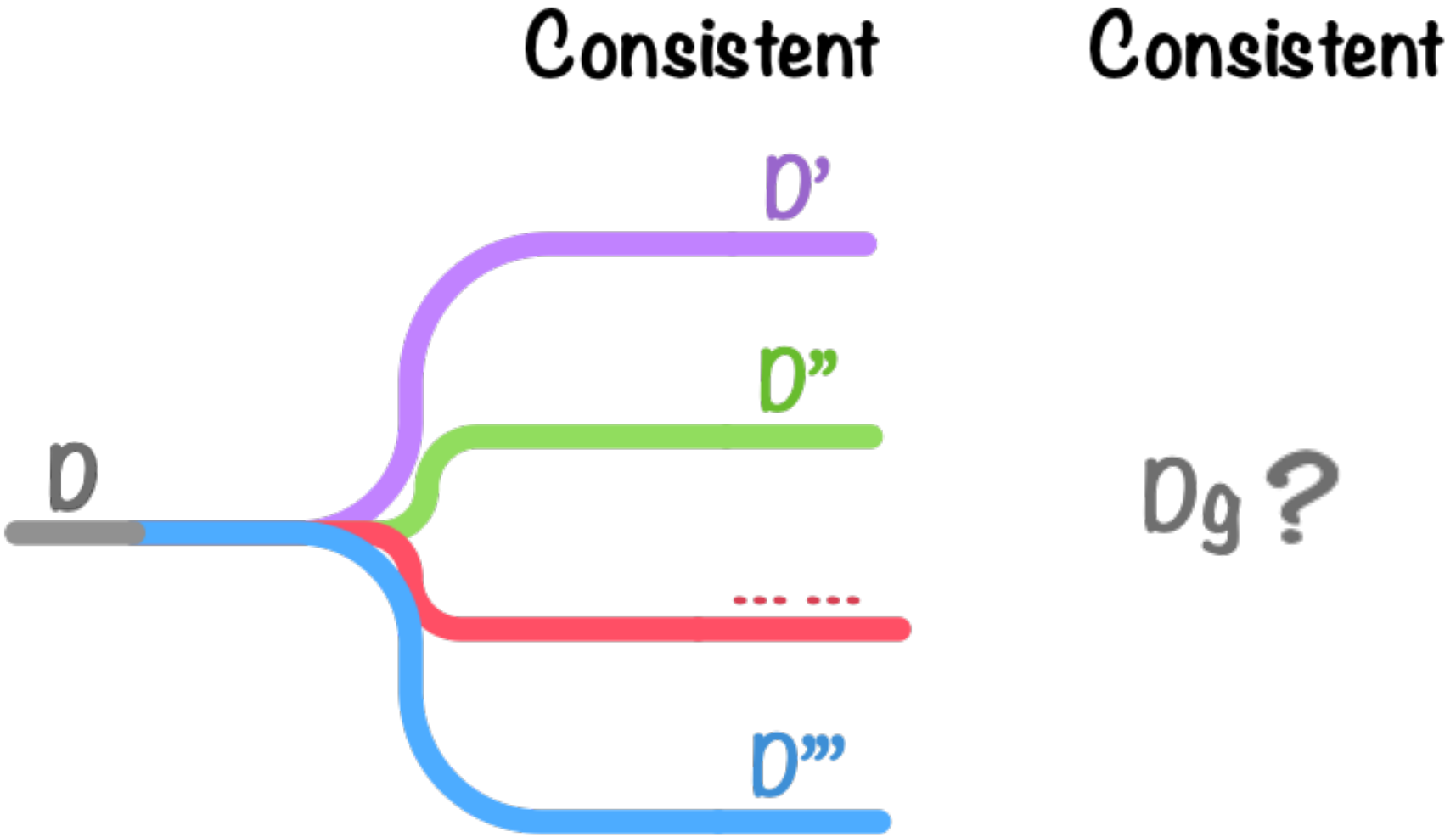
cap

	country	capital
s1	China	Beijing
s2	Canada	Ottawa
s3	…	…

CD:]t1, t2 (t1.salary > t2.salary and t1.tax < t2.tax)

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

COMPUTING A CONSISTENT DATABASE



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

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

CONFIDENCE VALUES INTERACTION

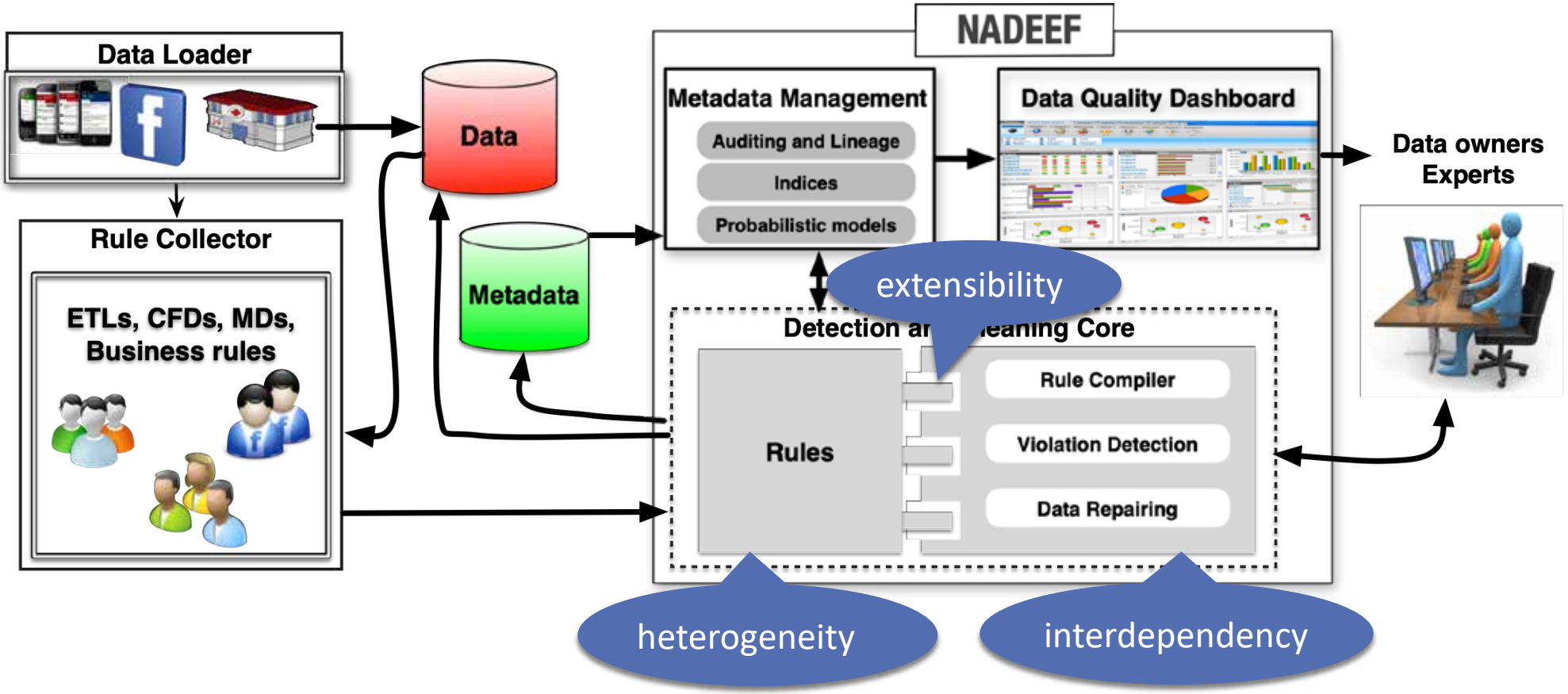
FD: [nationality] -> [capital]

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

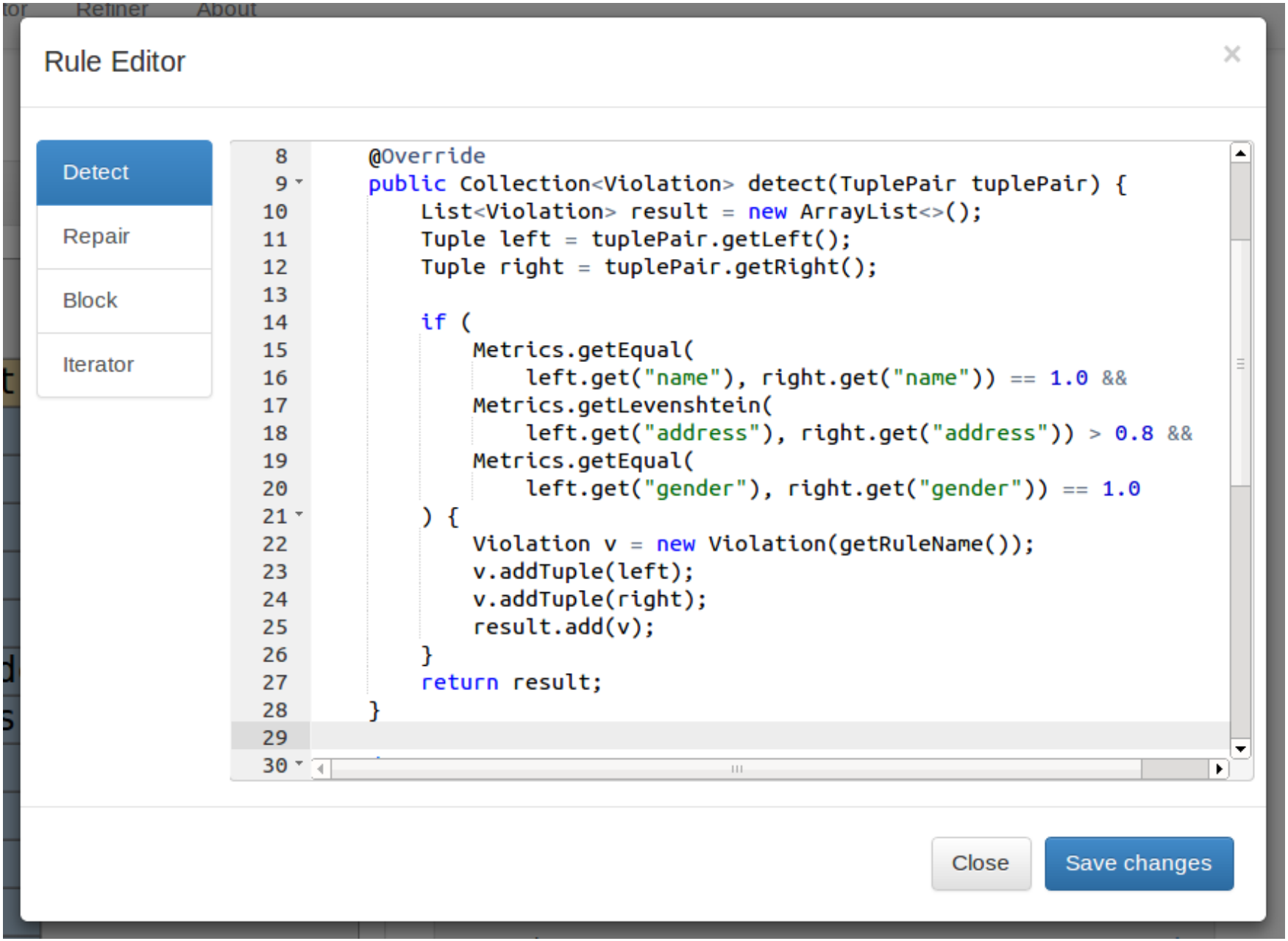
	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	Ottawa (1.0)	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

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

NADEEF


Data Loader
Rule Collector
ETLs, CFDs, MDs,
Business rules
Data
Metadata
NADEEF
Metadata Management
Auditing and Lineage
Indices
Probabilistic models
Data Quality Dashboard
extensibility
Rules
Rule Compiler
Violation Detection
Data Repairing
Data owners
Experts
heterogeneity
interdependency

NADEEF


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