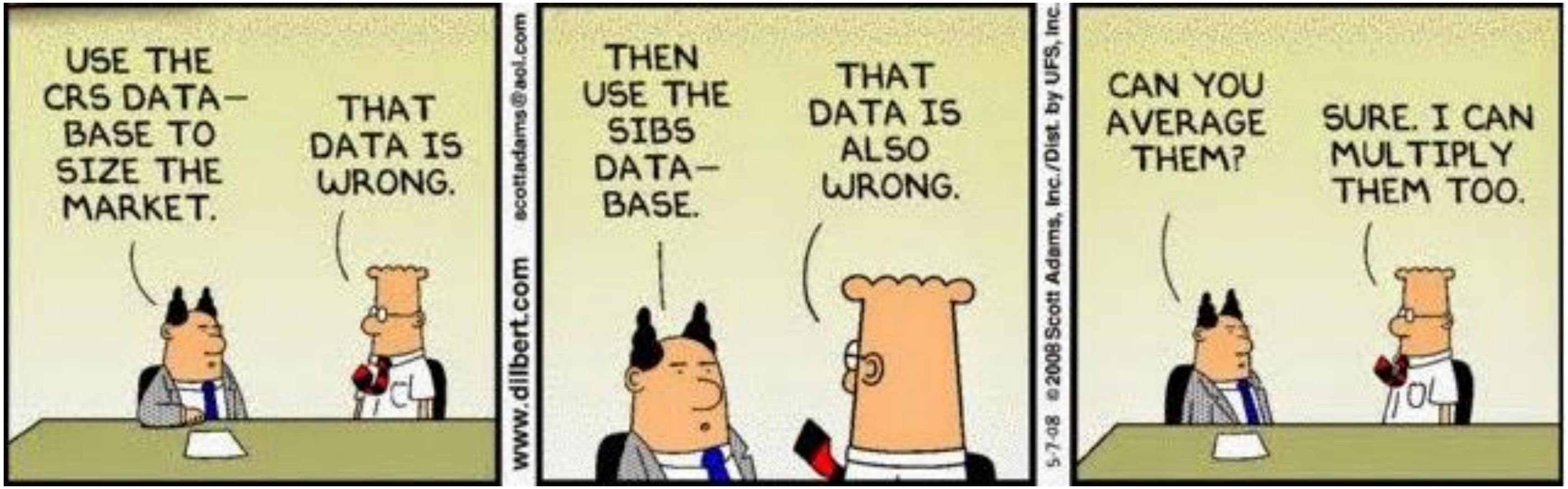


6.S079
Data Cleaning – Part 2

OUTLINE

Data Integration

- **Different schemas** →Schema matching
- **Duplicates** → Entity resolution
- **Contradicting data** → data fusion

Data Cleaning

- **Missing values** → Value imputation
- **Wrong data** → Outlier detection
- **Missing records** → Species estimation

WHY ARE THE VALUES MISSING?

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

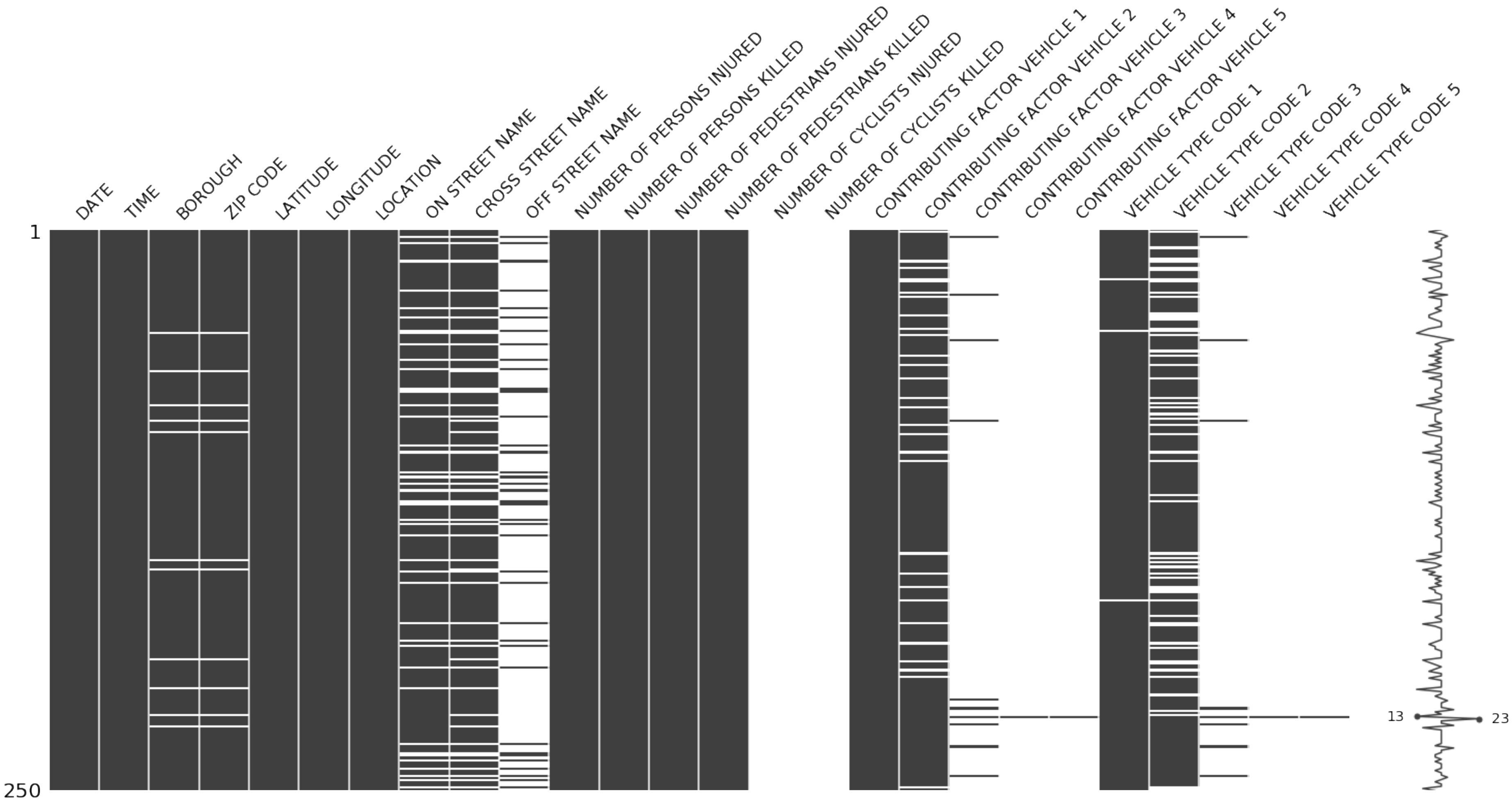
- **Missing at Random (MAR)**
  - Better name: Missing Conditionally at Random
  - Systematic relationship between the propensity of missing values and the *observed* data, but *not* the missing data.
    --> if we can control for this conditional variable, we can get a random subset.

- **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.

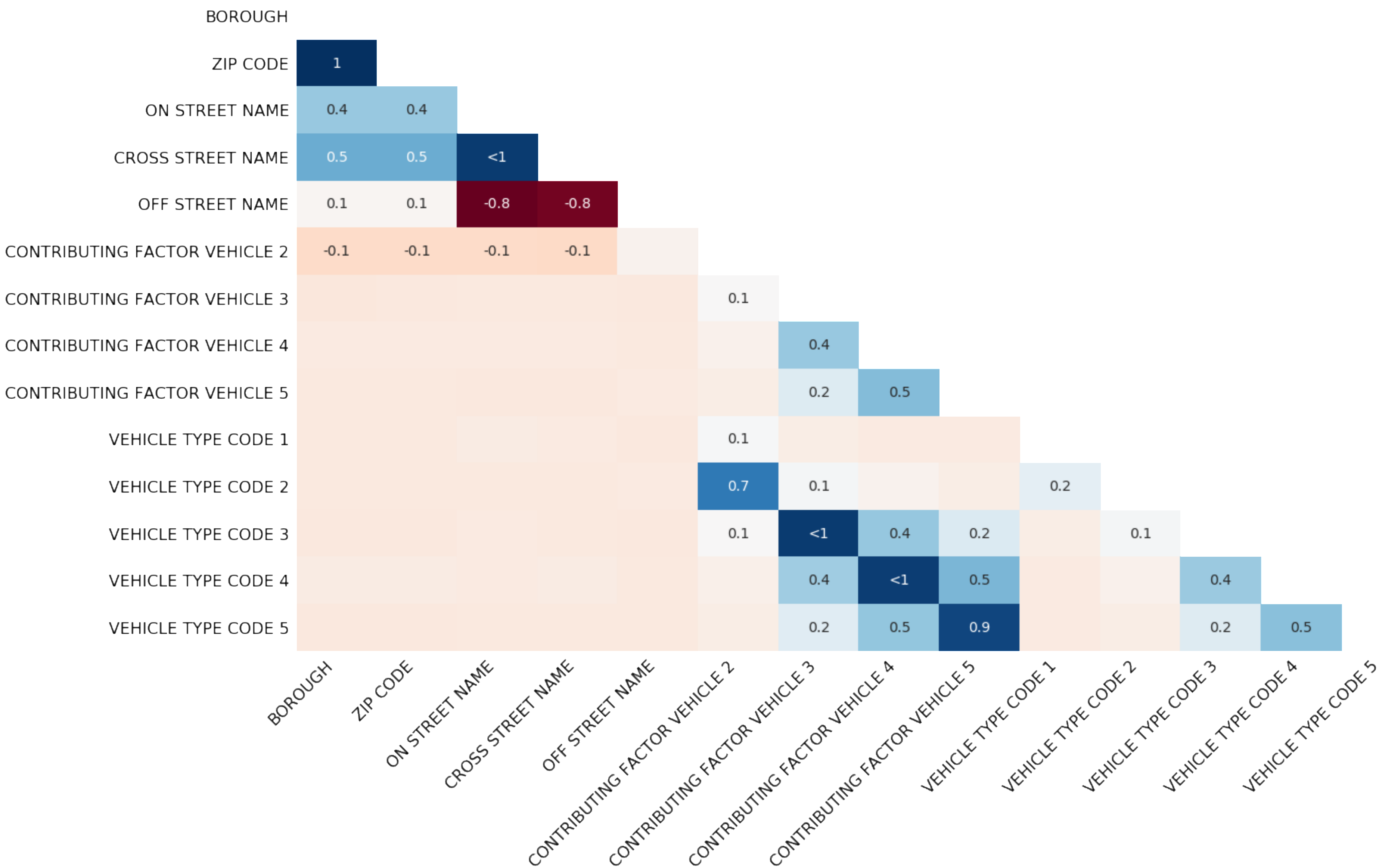
HOW DO YOU START ADDRESSING MISSING VALUES?

VISUALIZATIONS TO DETECT BIAS



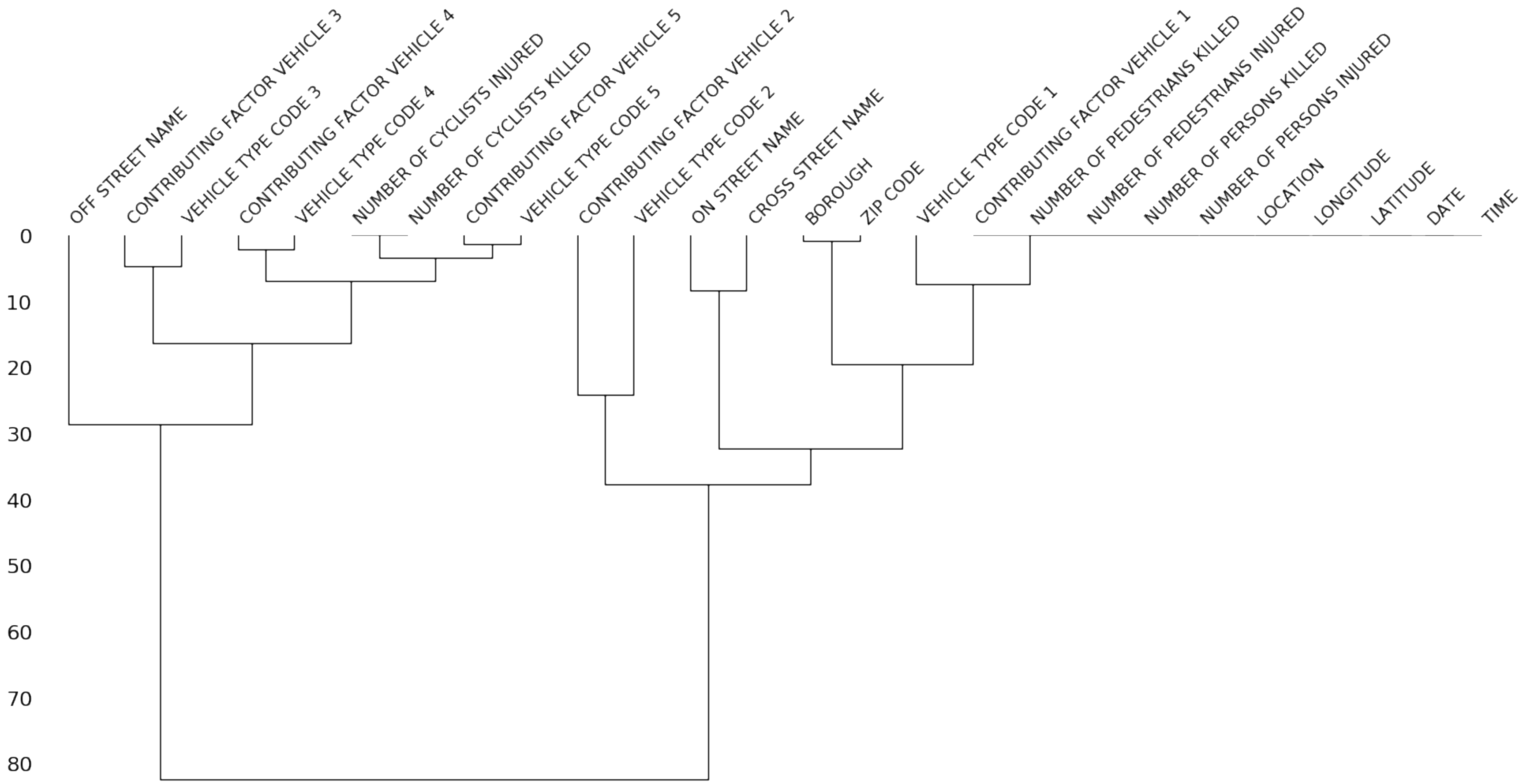
A lot of tips here: https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS



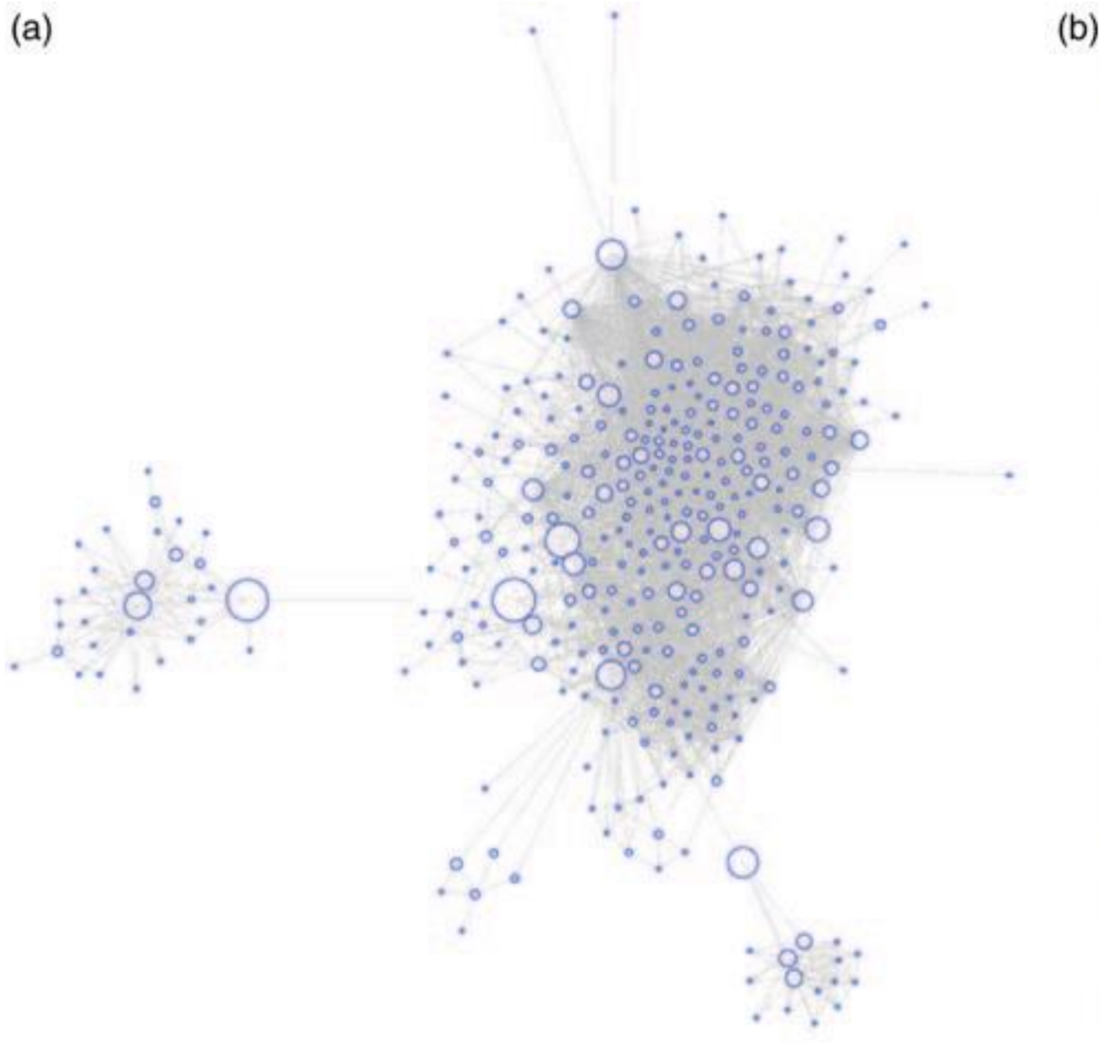
https://github.com/ResidentMario/missingno

VISUALIZATIONS TO DETECT BIAS



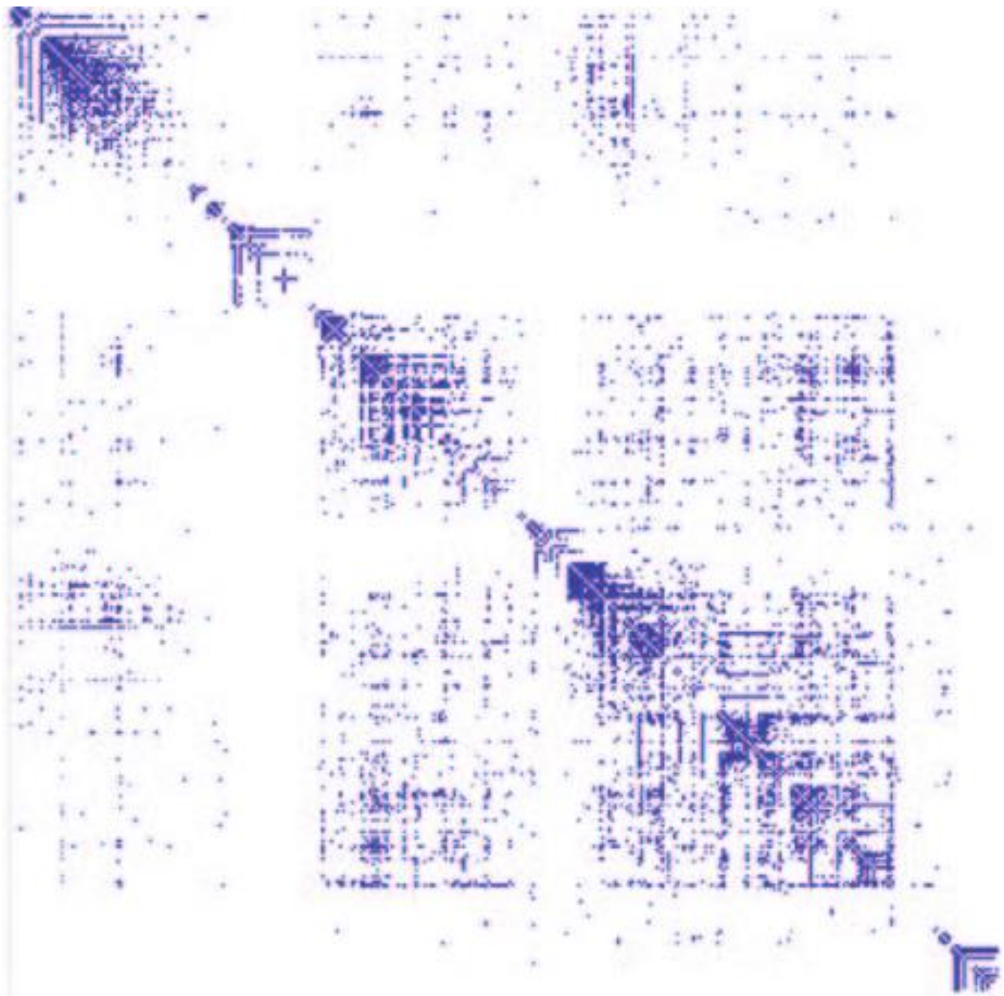
Alternative: Frequent pattern mining

FACEBOOK SOCIAL GRAPH: VISUALIZATION THE NODE-LINK DIAGRAM



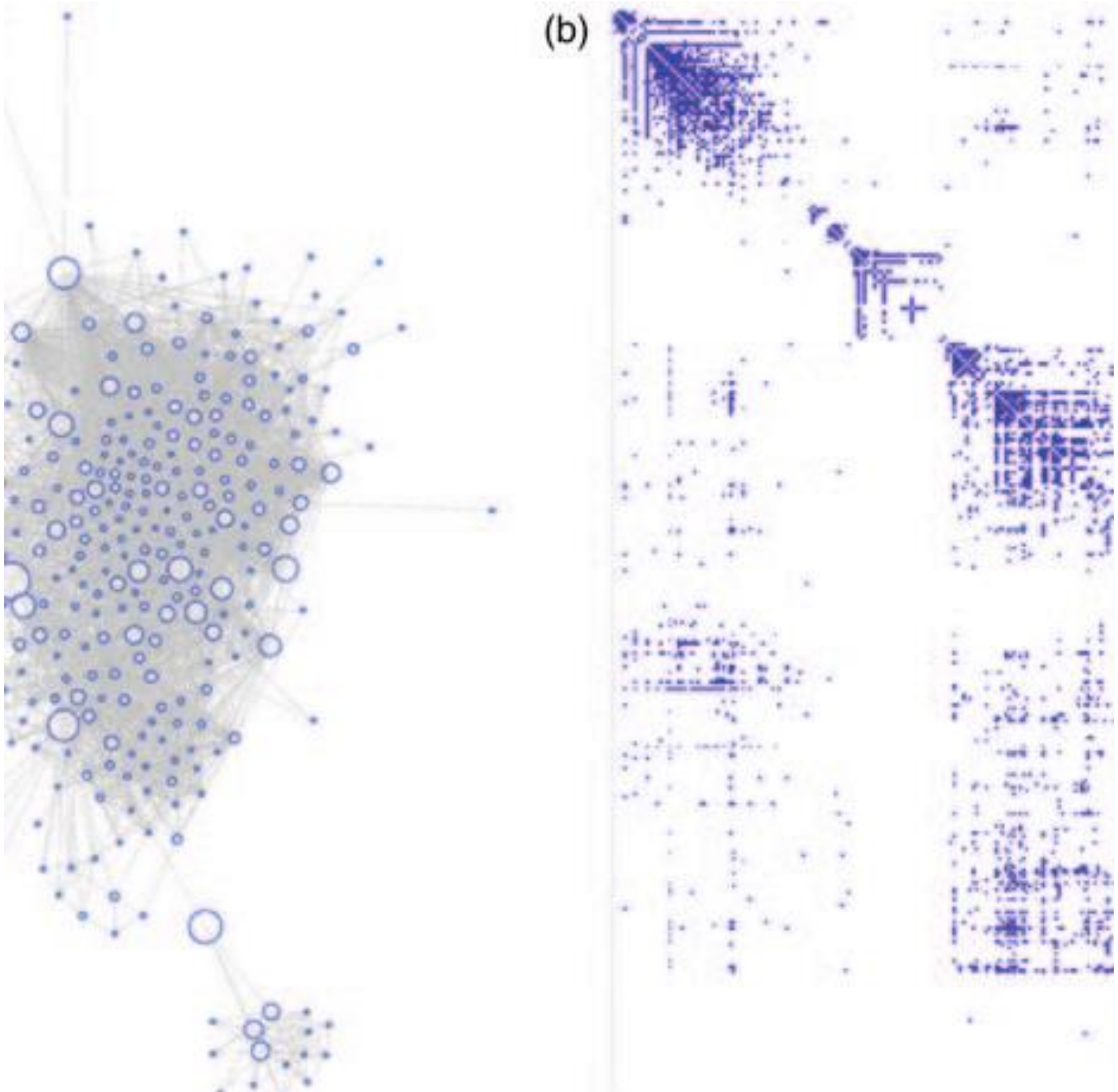
[Sean Kandel et al: Research directions in data wrangling: Visualizations and transformations for usable and credible data, Information Visualization, 2011]

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

CLASS EXERCISE

- **You are offered a new job as a SWE L4 at BOOBLE in the new Storage Division.**
- **They asked you to make a salary proposal before they make you an offer.**
- **Luckily, a year back some BOOBLE salary data got leaked and you are planning to use the average Base, Bonus, and Stock data to do a data-driven negotiation.**
- **How would you deal with the missing values to make an (1) unbiased/fair proposal and (2) a biased proposal to maximize your salary.**

BOOBLE salary data				
Role	**Devision**	**Base**	**Bonus**	**Stock**
SWE L4	Cloud	$ 150,000	$ 30,000	
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000
SWE L4	Ads	$ 160,000		
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000
SWE L4	Cloud		$ 20,000	$ 75,000
SWE L4	Cloud	$ 150,000		
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000

Total compensation = avg(base) + avg(bonus) + avg(stock)

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~~

SALARY EXAMPLE – PAIRWISE DELETION

BOOBLE salary data					
Role	Devision	Base	Bonus	Stock	
SWE L4	Cloud	$ 150,000	$ 30,000		
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000	
SWE L4	Ads	$ 160,000			
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000	
SWE L4	Cloud		$ 20,000	$ 75,000	
SWE L4	Cloud	$ 150,000			
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000	
Pairwise removel					
Sum		$ 975,000	$ 130,000	$ 333,000	
N		6	5	4	Total
AVG		$ 162,500	$ 26,000	$ 83,250	$ **271,750**

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~~

SALARY EXAMPLE – LISTWISE DELETION

BOOBLE salary data					
Role	Devision	Base	Bonus	Stock	
SWE L4	Cloud	$ 150,000	$ 30,000		
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000	
SWE L4	Ads	$ 160,000			
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000	
SWE L4	Cloud		$ 20,000	$ 75,000	
SWE L4	Cloud	$ 150,000			
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000	
Pairwise removel					
Sum		$ 975,000	$ 130,000	$ 333,000	
N		6	5	4	Total
AVG		$ 162,500	$ 26,000	$ 83,250	$ **271,750**

Listwise removal					
Role	Devision	Base	Bonus	Stock	
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000	
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000	
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000	**Total**
AVG		$ 171,667	$ 26,667	$ 86,000	**$ 284,333**

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

- 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)

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

MULTIVARIATE 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.

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

MULTIVARIATE IMPUTATION

Regression imputation

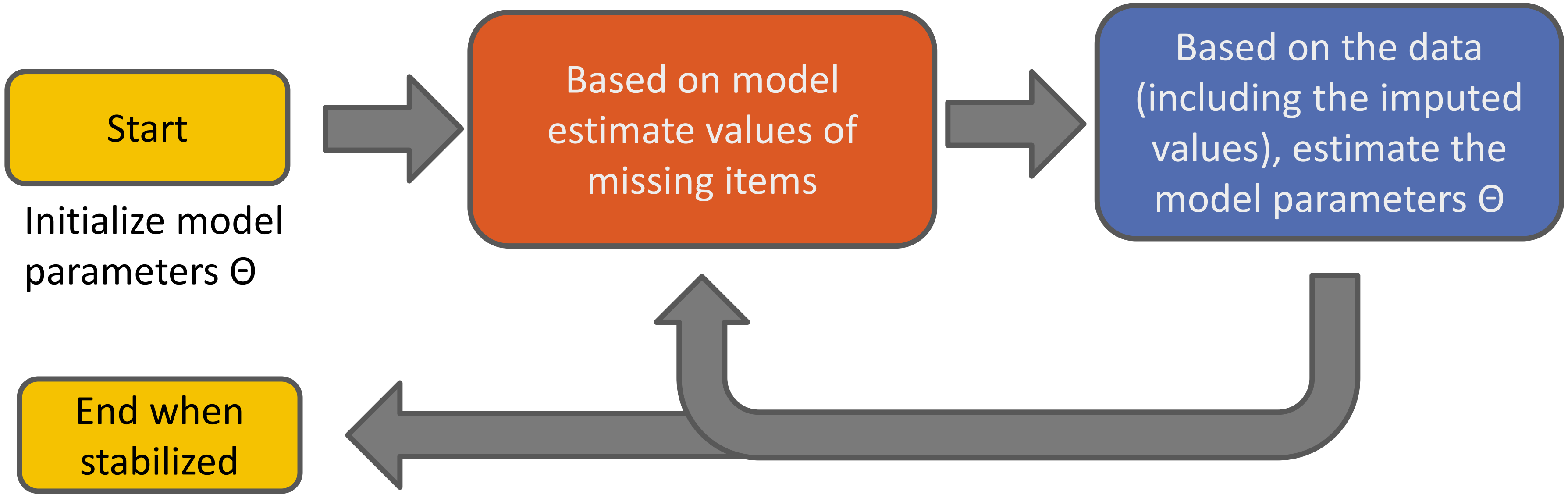
- Replace missing values with predicted score from regression equation. Use complete cases to regress the variable with incomplete data on the other complete variables.
- Advantage: Uses information from the observed data, gives better results than previous ones
- Disadvantage: over-estimates model fit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

Maximum Likelihood Estimation (MICE)

- Identifies the set of parameter values that produces the highest log-likelihood.
- ML estimate: value that is most likely to have resulted in the observed data

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

EM ALGORITHM



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.

We will cover this algorithm in more depth later

MULTIVARIATE SINGLE IMPUTATION

Nearest neighbors imputation

- KNN defines for each sample or individual a set of K-nearest neighbors and then replaces the missing data for a given variable by averaging (non-missing) values of its neighbors
- Advantage: Simple, uses information from the observed data, experimentally shows good performance
- Disadvantage: not statistically grounded, might over-estimates model fit and correlation

Fuzzy K-means Clustering

Bayesian Principal Component Analysis

Deep Learning-based approaches

….

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

SALARY EXAMPLE - 1NN

SWE L4	Cloud	$ 150,000.00	$ 30,000.00	$ 75,000.00	
SWE L4	Brain	$ 170,000.00	$ 25,000.00	$ 80,000.00	
SWE L4	Ads	$ 160,000.00	$ 20,000.00	$ 78,000.00	
SWE L4	Brain	$ 185,000.00	$ 35,000.00	$ 100,000.00	
SWE L4	Cloud	$ 160,000.00	$ 20,000.00	$ 75,000.00	
SWE L4	Cloud	$ 150,000.00	$ 30,000.00	$ 75,000.00	
SWE L4	Cloud	$ 160,000.00	$ 20,000.00	$ 78,000.00	**Total**
AVG		**$ 162,142.86**	**$ 25,714.29**	**$ 80,142.86**	**$ 268,000**

Pairwise removal

BOOBLE salary data					
Role	Devision	Base	Bonus	Stock	
SWE L4	Cloud	$ 150,000	$ 30,000		
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000	
SWE L4	Ads	$ 160,000			
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000	
SWE L4	Cloud		$ 20,000	$ 75,000	
SWE L4	Cloud	$ 150,000			
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000	
AVG		**$ 162,500**	**$ 26,000**	**$ 83,250**	$ **271,750**

Listwise removal

Role	Devision	Base	Bonus	Stock	
SWE L4	Brain	$ 170,000	$ 25,000	$ 80,000	
SWE L4	Brain	$ 185,000	$ 35,000	$ 100,000	
SWE L4	Cloud	$ 160,000	$ 20,000	$ 78,000	**Total**
AVG		**$ 171,667**	**$ 26,667**	**$ 86,000**	**$ 284,333**

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 max($n_{missing}$) 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 $n\text{-}n_{missing}$, 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
- 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

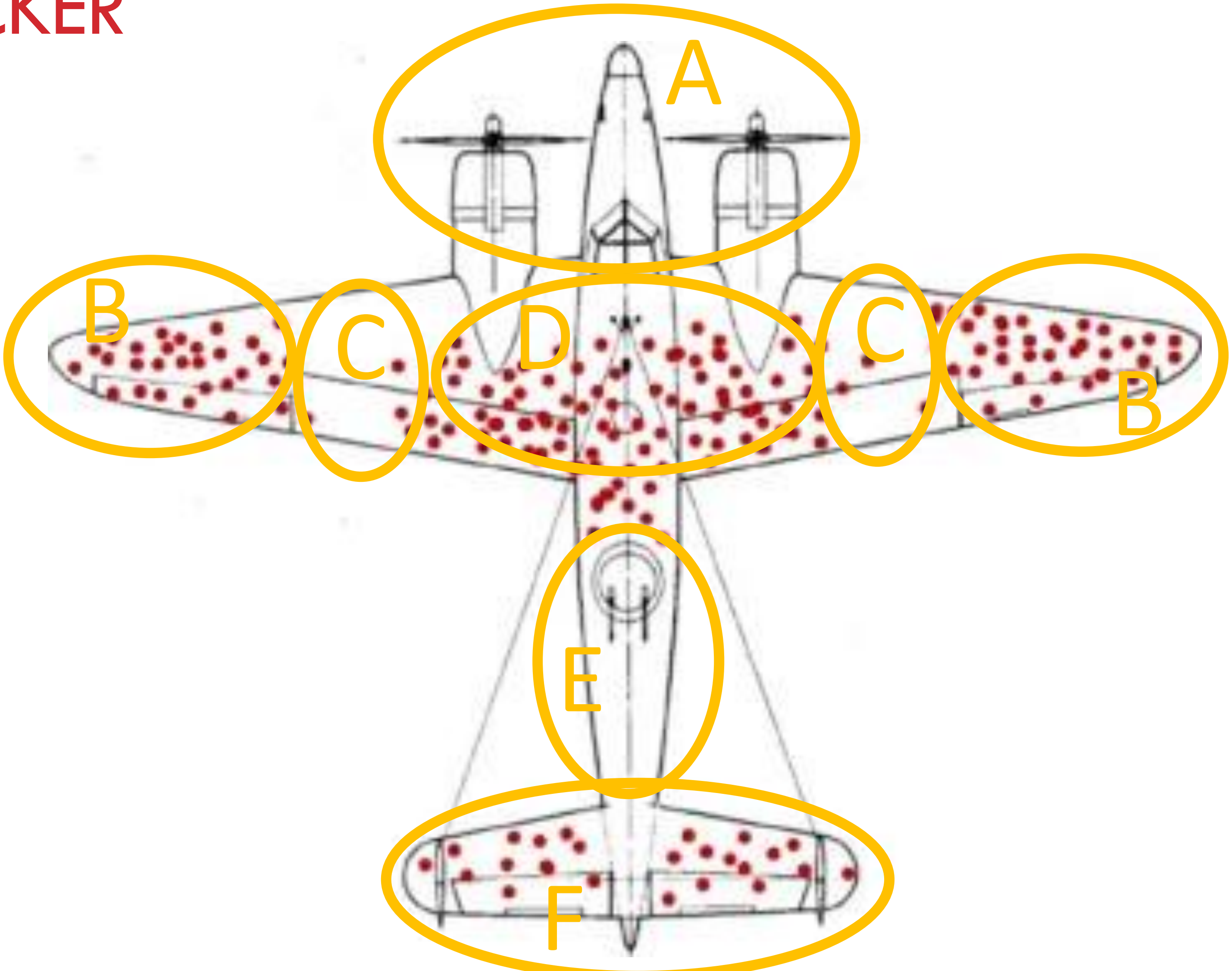
Name	Address	#Employees	Revenue (M)	Profit (M)
Google	1600 Amphitheatre Parkway, Mountain View, CA	50000	$100000M	$40000M
Apple	1 Infinite Loop; Cupertino, CA 95014, USA	70000	$200000M	$50000M
IBM	1 New Orchard Rd; 10504	400000	$100000M	null
Microsoft	Albuquerque, New Mexico	130000	$125000M	$40000M
Tableau	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

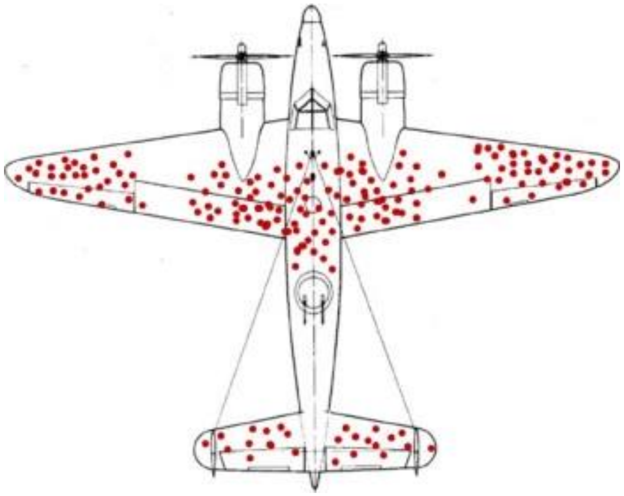


Where would you enforce the plane?

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



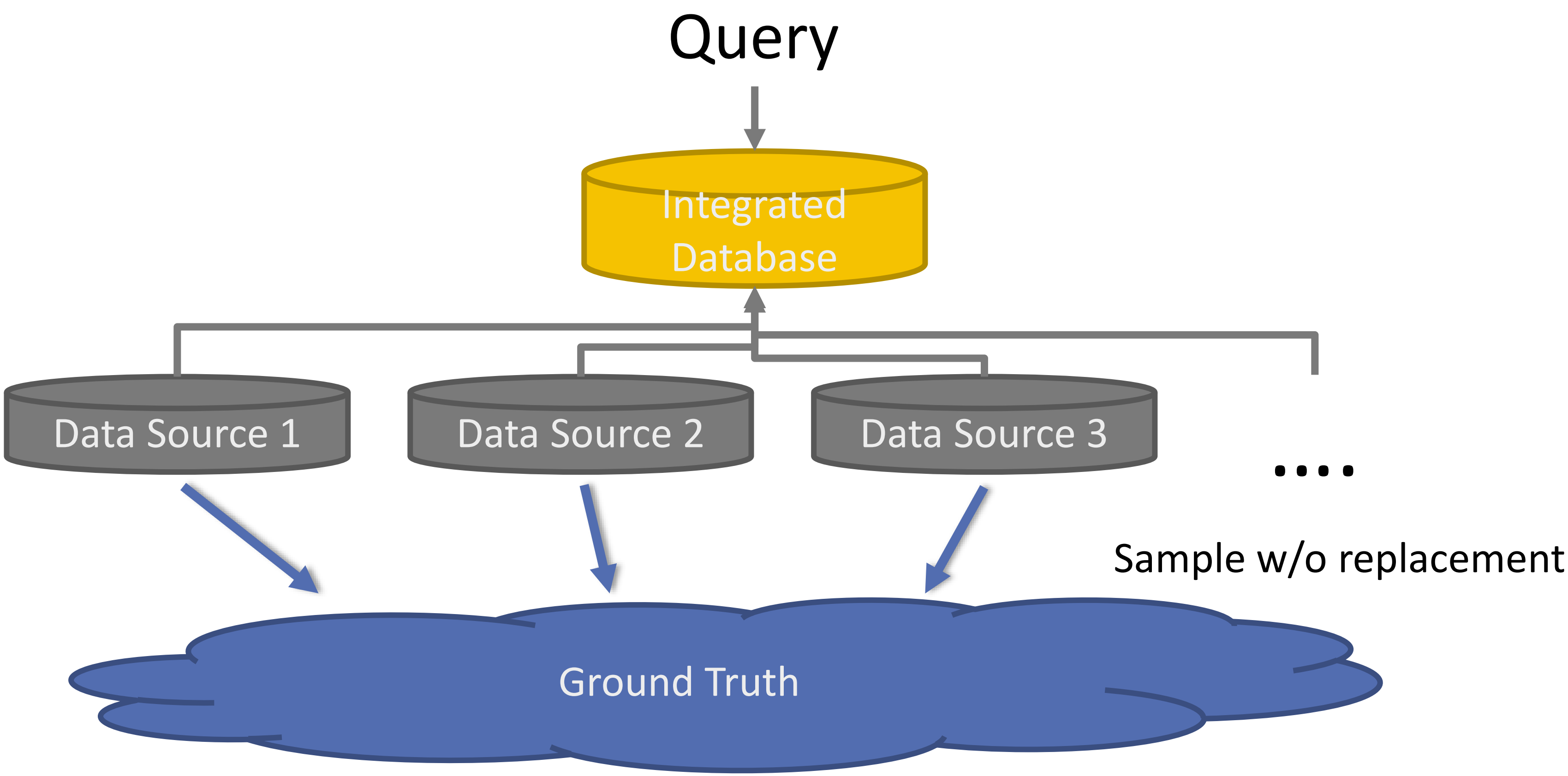
VS

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

f_2: 1

f_4: 2

f_5: 2

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

$$\hat{N} = \frac{\boxed{c}}{\left(1 - \boxed{f_1/_n}\right)}$$

Missing mass

Number of Unknown Unknowns:

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

*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
```

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

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

M: Estimate of Unknown Unknowns Count

ϕ: 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

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

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

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

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

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
….	…	..	…	…	…	…
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	[illegible] United States	20	null	[illegible]

Encoding Error
(nb in thousands)

Rule Violations

Outdated data / wrong data

Spelling mistakes / abbreviations

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

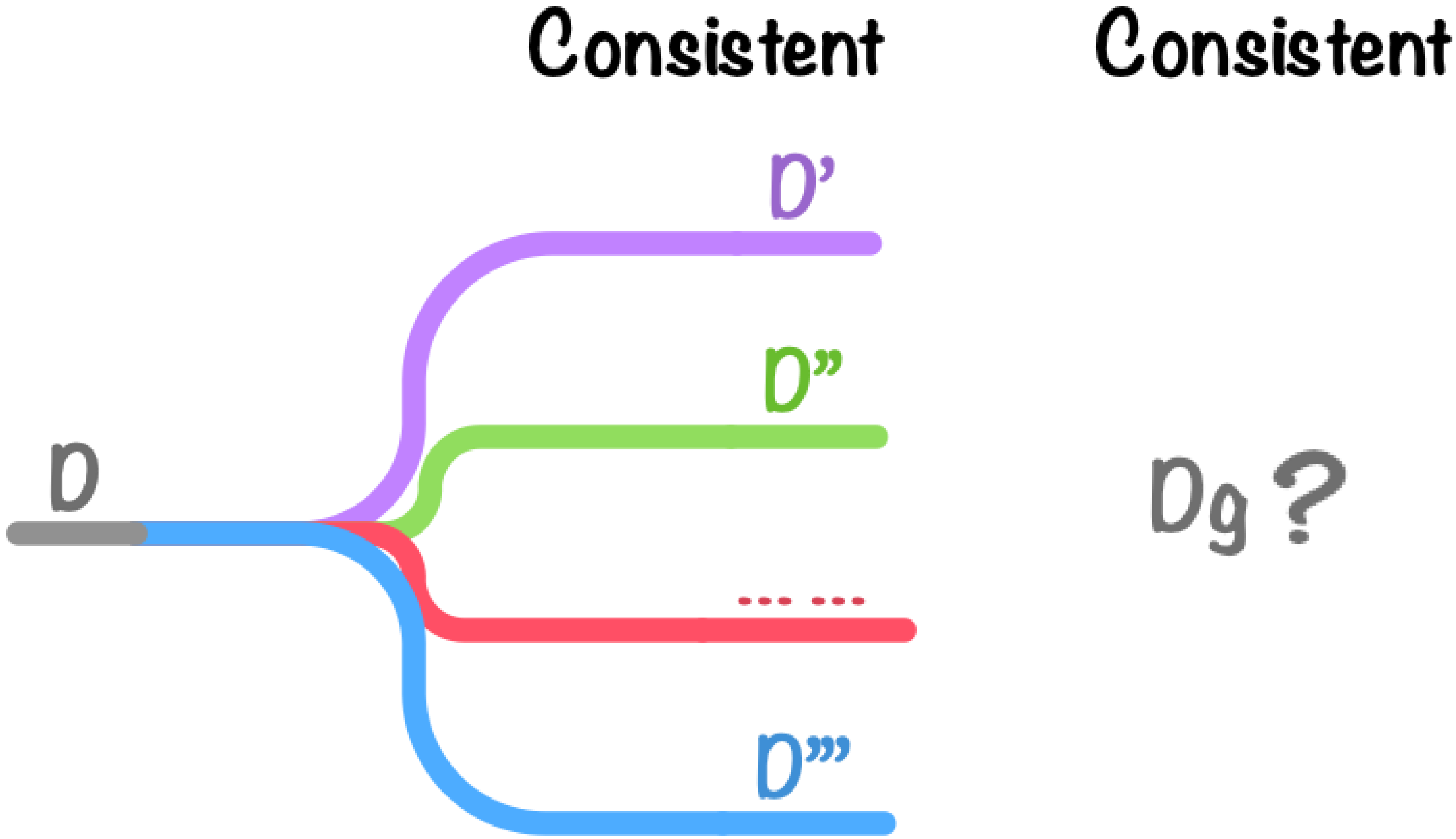
cap

	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

Equivalence class

Vertex cover

SAT solver

… …

CONFIDENCE VALUES INTERACTION

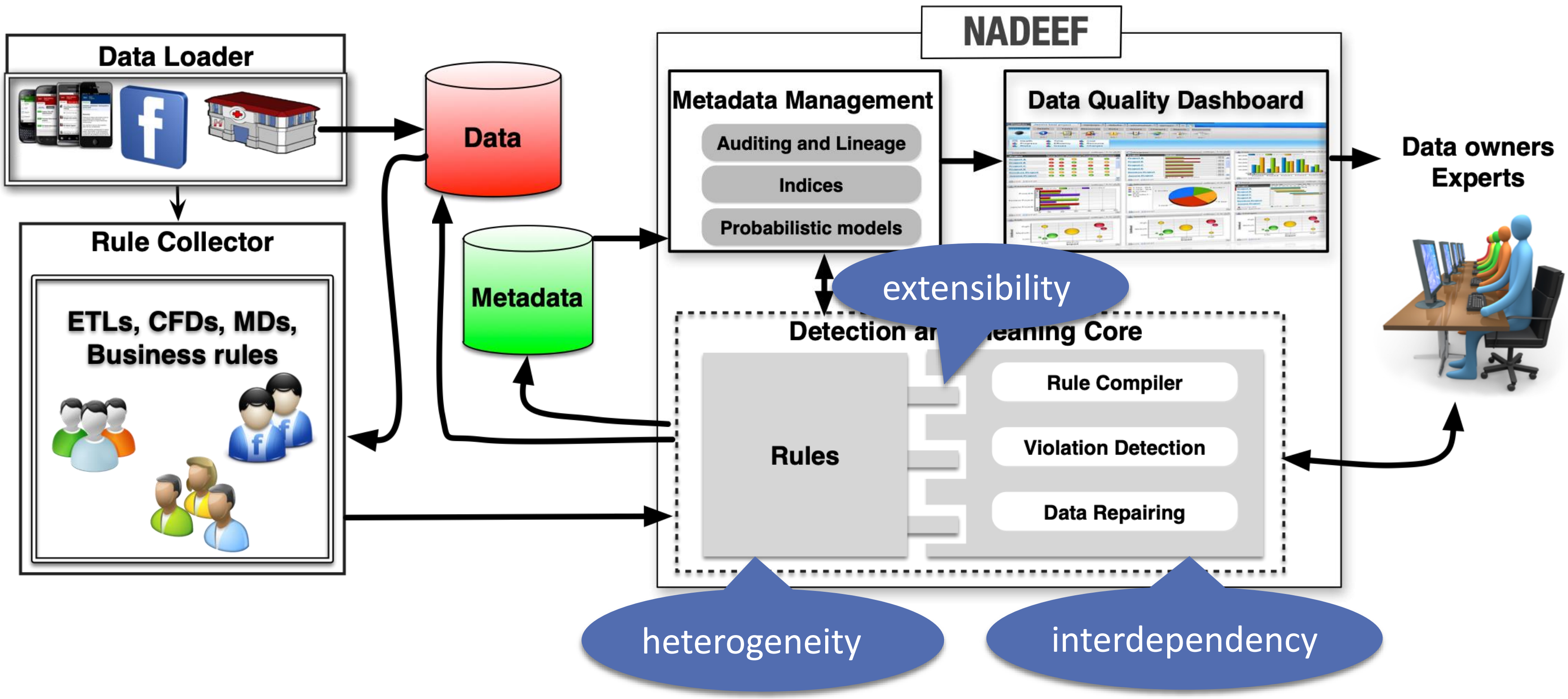
FD: [nationality] -> [capital]

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

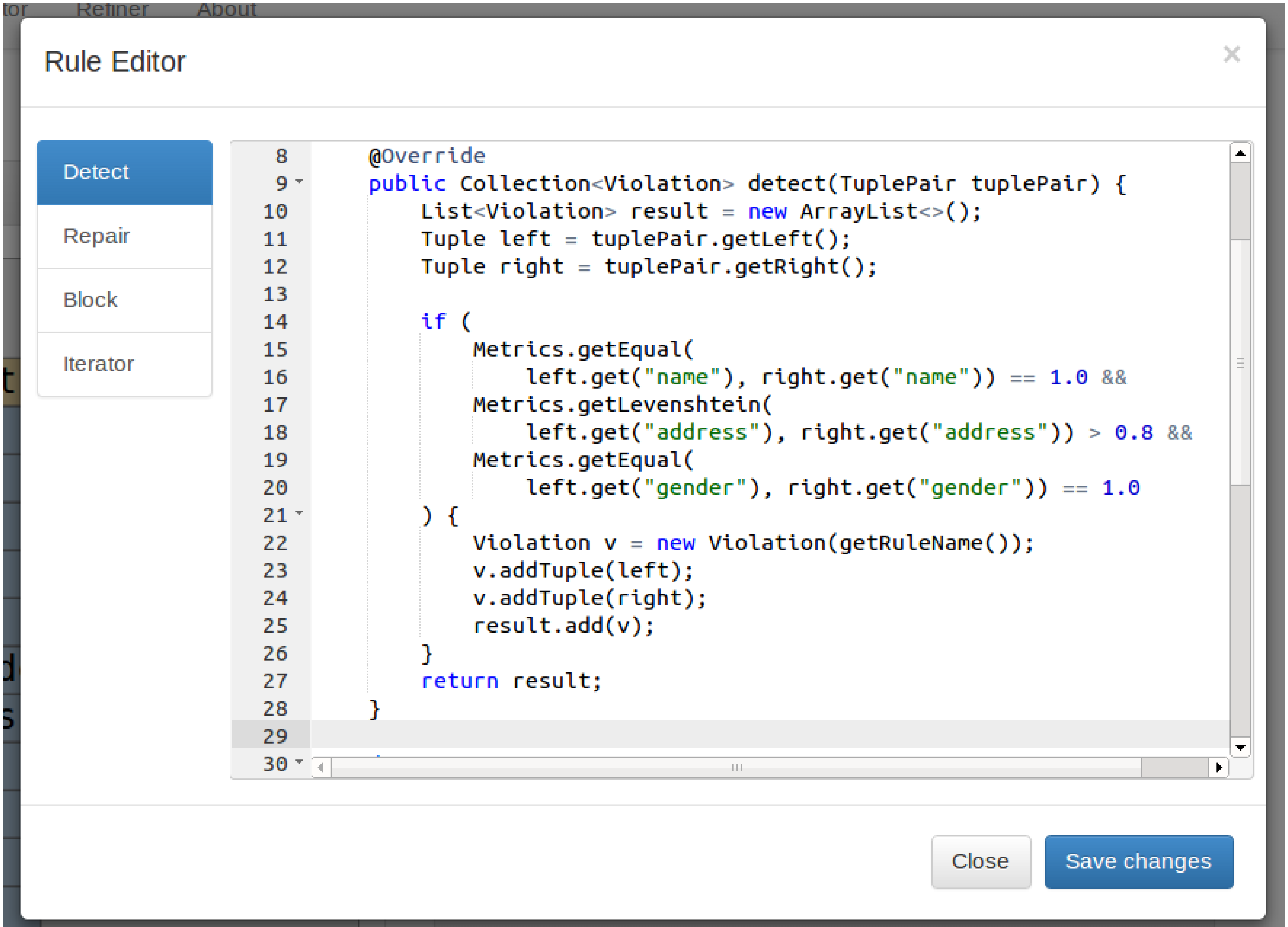
	name	nationality	capital	bornAt
r1	Nan (0.9)	China (1.0)	Beijing (1.0)	Shenyang (0.9)
r2	Yan (0.8)	China (1.0)	Beijing (0.5)	Hangzhou (0.9)
r3	Si (0.9)	Canada (1.0)	Ottawa (1.0)	Changsha (0.8)
r4	Miura (0.9)	Canada (0.9)	Vancuver (0.5)	Kyoto (1.0)

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

NADEEF


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

NADEEF



tor
Refiner
About
Rule Editor
Detect
Repair
Block
Iterator
@Override
public Collection<Violation> detect(TuplePair tuplePair) {
List<Violation> result = new ArrayList<>();
Tuple left = tuplePair.getLeft();
Tuple right = tuplePair.getRight();
if (
Metrics.getEqual(
left.get("name"), right.get("name")) == 1.0 &&
Metrics.getLevenshtein(
left.get("address"), right.get("address")) > 0.8 &&
Metrics.getEqual(
left.get("gender"), right.get("gender")) == 1.0
) {
Violation v = new Violation(getRuleName());
v.addTuple(left);
v.addTuple(right);
result.add(v);
}
return result;
}
Close
Save changes

OUTLIER DETECTION

ANOMALY/OUTLIER DETECTION

What are anomalies/outliers?

- The set of data points that are considerably different than the remainder of the data

Variants of Anomaly/Outlier Detection Problems

- Given a database D, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
- Given a database D, find all the data points $\mathbf{x} \in D$ having the top-n largest anomaly scores $f(\mathbf{x})$
- Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

Applications:

- Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

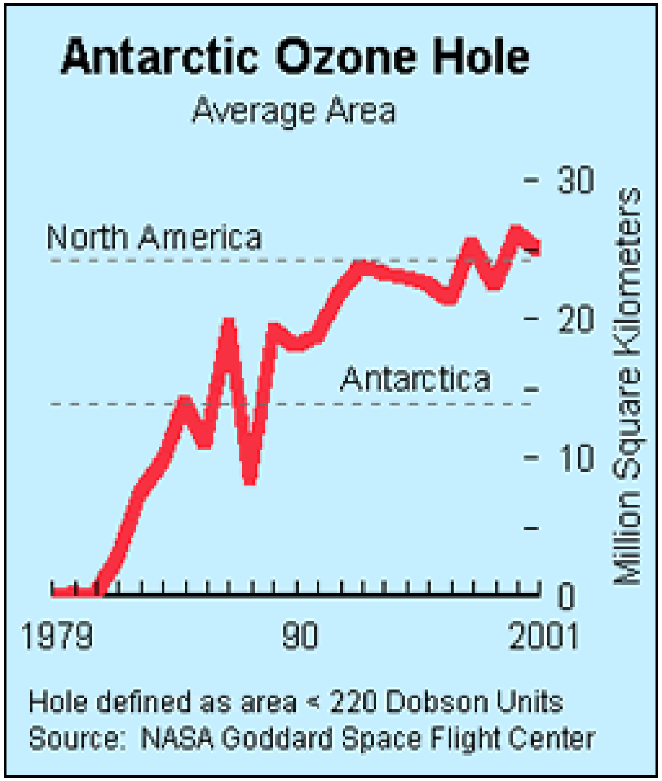
IMPORTANCE OF ANOMALY DETECTION

Ozone Depletion History

In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

ANOMALY DETECTION

Challenges

- How many outliers are there in the data?
- Method is unsupervised
  - Validation can be quite challenging (just like for clustering)
- Finding needle in a haystack

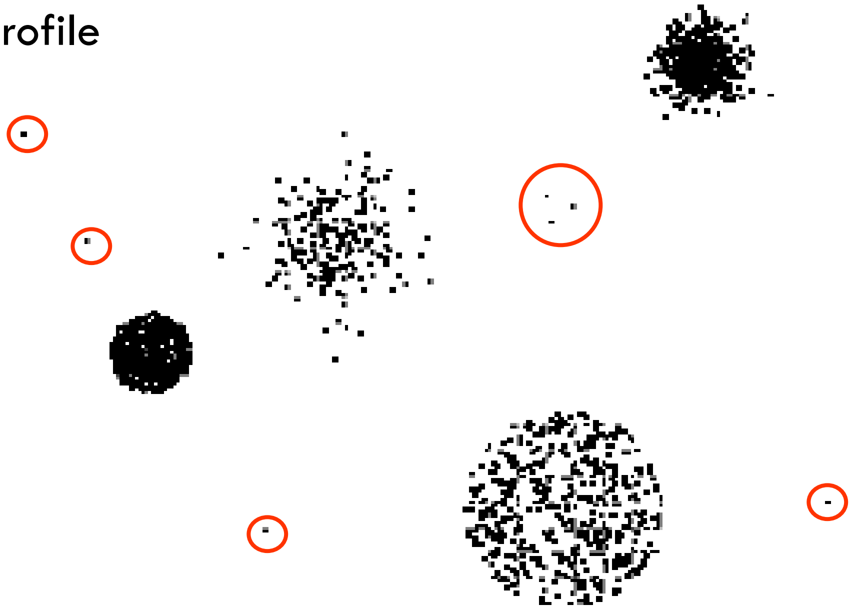
Working assumption:

- There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data

ANOMALY DETECTION SCHEMES

General Steps

- Build a profile of the "normal" behavior
  - Profile can be patterns or summary statistics for the overall population
- Use the "normal" profile to detect anomalies
  - Anomalies are observations whose characteristics differ significantly from the normal profile



Types of anomaly detection schemes

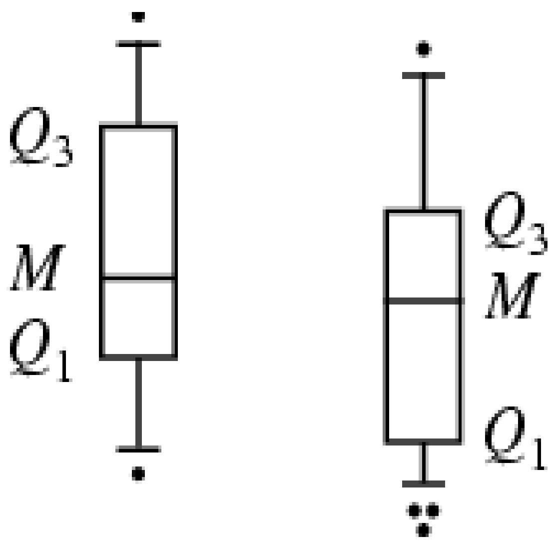
- Graphical
- Model-based
- Distance-based
- Clustering-based

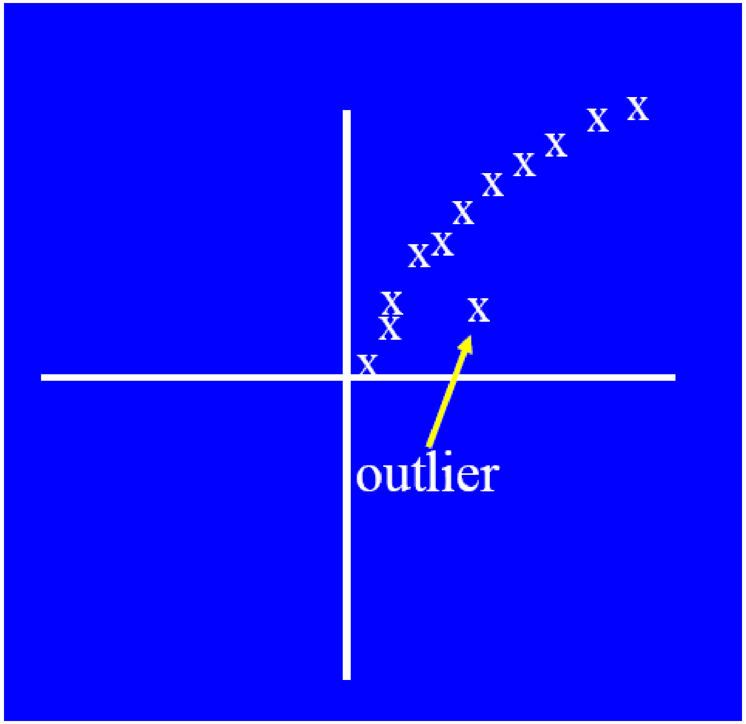
GRAPHICAL APPROACHES

Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

Limitations

- Time consuming
- Subjective



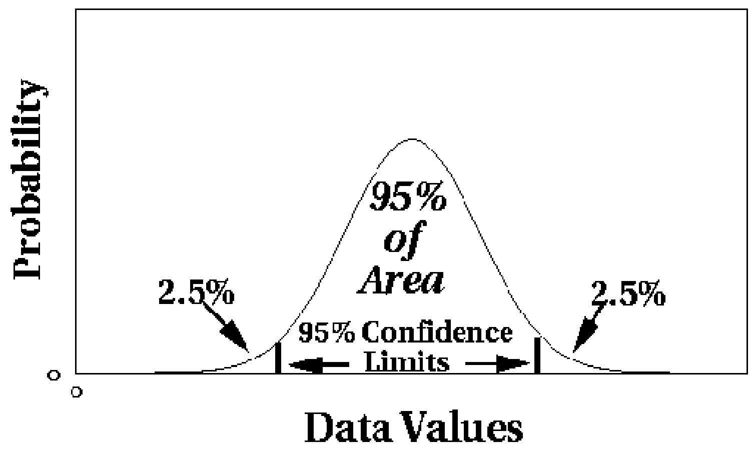


STATISTICAL APPROACHES---MODEL-BASED

Assume a parametric model describing the distribution of the data (e.g., normal distribution)

Apply a statistical test that depends on

- Data distribution
- Parameter of distribution (e.g., mean, variance)
- Number of expected outliers (confidence limit)



GRUBBS' TEST

Detect outliers in univariate data

Assume data comes from normal distribution

Detects one outlier at a time, remove the outlier, and repeat

- H_0: There is no outlier in data
- H_A: There is at least one outlier

Grubbs' test statistic:

Reject H_0 if:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

$$G > \frac{(N-1)}{\sqrt{N}}\sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

STATISTICAL-BASED – LIKELIHOOD APPROACH

Assume the data set D contains samples from a mixture of two probability distributions:

- M (majority distribution)
- A (anomalous distribution)

General Approach:

- Initially, assume all the data points belong to M
- Let $L_t(D)$ be the log likelihood of D at time t
- For each point x_t that belongs to M, move it to A
  - Let L_{t+1} (D) be the new log likelihood.
  - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
  - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

LIMITATIONS OF STATISTICAL APPROACHES

Most of the tests are for a single attribute

In many cases, data distribution/model may not be known

For high dimensional data, it may be difficult to estimate the true distribution

DISTANCE-BASED APPROACHES

Data is represented as a vector of features

Three major approaches

- Nearest-neighbor based
- Density based
- Clustering based

NEAREST-NEIGHBOR BASED APPROACH

Approach:

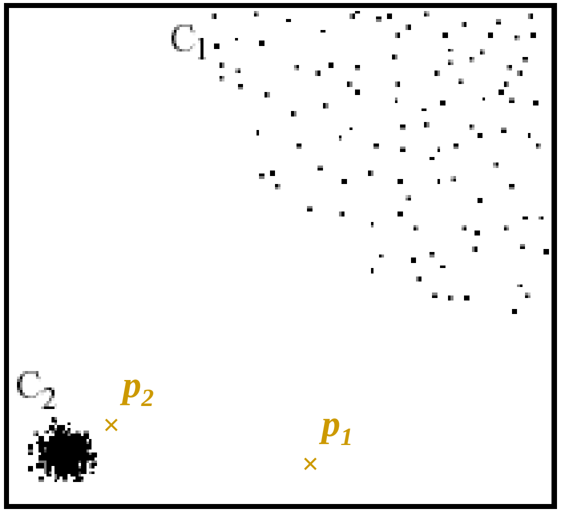
- Compute the distance between every pair of data points
- There are various ways to define outliers:
  - Data points for which there are fewer than *p* neighboring points within a distance *D*
  - The top n data points whose distance to the kth nearest neighbor is greatest
  - The top n data points whose average distance to the k nearest neighbors is greatest

DENSITY-BASED: LOF APPROACH

For each point, compute the density of its local neighborhood; e.g. use DBSCAN's approach

Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors

Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Alternative approach: directly use density function; e.g. DENCLUE's density function

CLUSTERING-BASED

Idea: Use a clustering algorithm that has some notion of outliers!

Problem what parameters should I choose for the algorithm; e.g. DBSCAN?

Rule of Thumb: Less than x% of the data should be outliers (with x typically chosen between 0.1 and 10); x might be determined with other methods; e.g. statistical tests.

	FN	LN	St	city	CC	country	tel	gd
t_1:	David	Jordan	12 Holywell Street	Oxford	44	UK	66700543	Male
t_2:	Paul	Simon	5 Ratcliffe Terrace	Oxford	44	UK	44944631	Male

(a) D_1: An instance of schema bank

	FN	LN	str	city	CC	country	phn	when	where
r_1:	David	Jordan	12 Holywell Street	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
r_2:	Paul	Simon	5 Ratcliffe Terrace	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
r_3:	David	Jordan	12 Holywell Street	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
r_4:	Peter	Austin	7 Market Street	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

(b) Database D_2: An instance of schema tran

r1: (on table tran) if a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands;

r2: (on tables bank and tran) if the same person from different tables has different phones, the phone number from table bank is more reliable;

r3: (on table tran) a country code (CC) uniquely determines a country;

r4: (on table tran) if two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour (assuming 6 hours' time difference between these two countries), these two purchases are either a fraud or were erroneously recorded.

```
Class Rule1 {                                                    /* for φ1 */
  set⟨cell⟩ vio(Tuple s1) {                              /*s1 in table tran */
    if (s1[CC]=31 ∧ (s1[country] ≠Netherlands ∨ s1[country] ≠Holland))
      return { s1[CC, country]; }
    return ∅;
  }
  set⟨Expression⟩ fix (set⟨cell⟩ V) {
    return { V.s[country] ← Netherlands; }
  } } /* end of class definition */

Class Rule2 {                                                    /* for φ2 */
  set⟨cell⟩ vio (Tuple s1, Tuple s2) {            /*s1 in bank, s2 in tran */
    if (s1[LN, St, city]=s2[LN, str, city]∧s1[FN] ≈ s2[FN]∧s1[tel]≠s2[phn])
      return { s1[FN, LN, St, city, tel], s2[FN, LN, str, city, phn]; }
    return ∅;
  }
  set⟨Expression⟩ fix (set⟨cell⟩ V) {
    return { V.s2[phn] ← V.s1[tel]; }
  } } /* end of class definition */

Class Rule3 {                                                    /* for φ3 */
  set⟨cell⟩ vio (Tuple s1, Tuple t2) {              /* s1, s2 in table tran */
    if (s1[CC] = s2[CC] ∧ s1[country] ≠ s2[country])
      return { s1[CC, country], s2[CC, country]; }
    return ∅;
  }
  set⟨Expression⟩ fix (set⟨cell⟩ V) {
    set⟨Expression⟩ fixes;
    fixes.insert(V.s1[country] ←V.s2[country]);
    fixes.insert(V.s2[country] ←V.s1[country]);
    return fixes;
  } } /* end of class definition */

Class Rule4 {                                                    /* for φ4 */
  set⟨cell⟩ vio (Tuple s1, Tuple s2) {               /*s1, s2 in table tran */
    if (s1[LN, city, CC, tel] = s2[LN, city, CC, tel]
      ∧s1[where] = Netherlands ∧ s2[where] = US ∧ s1[FN] ≈ s2[FN]
      ∧(s1[when] − s2[when] ≥ 5) ∧ (s1[when] − s2[when] ≤ 7)
      return { s1[FN, LN, city, CC, tel, when, where],
        s2[FN, LN, city, CC, tel, when, where]; }
    return ∅;
  } } /* end of class definition */
```

Figure 3: Sample rules

WHY IS FINDING VIOLATIONS EXPENSIVE?