



Lecture 08: Dealing with missing values

Handling missing values

- Deletion
 - Delete the rows or columns with missing value
- Imputation
 - Replace cells containing missing values with something meaningful
- Imputation is good. But we cannot always impute
 - MNAR data should never be imputed

Types of Missing values

- Missing completely at random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

	Observed data	Unobserved data	Examples
MCAR	No relation	No relation	A random student's answer sheet is missing. Every answer sheet has equal probability of missing
MAR	Related	No relation	A student's answer sheet missing because he was absent
MNAR	No relation	Related	A student did not answer a question because he did not study that part

- Simple
- Reduces sample size
- Less degrees of freedom
- Information loss

Handling missing values

- Maintains sample size
- Introduces Bias

Deletion

Delete rows

Delete columns

Imputation

Simple (Univariate)

Multivariate

Non-IID data

IID data

IID data

Forward/
backward
filling

Constant

Grouped
Mean

kNN

Iterative

Linear
interpolation

Mean
Median
Mode

MICE

- Imputation only for MCAR/MAR data
- MNAR data has to be deleted

Perils of constant imputation

- Better to create a unknown unmapped feature to investigate separately



Mean Imputation

Price
100
90
50
40
20
100
60
120
200

Mean = 86.66

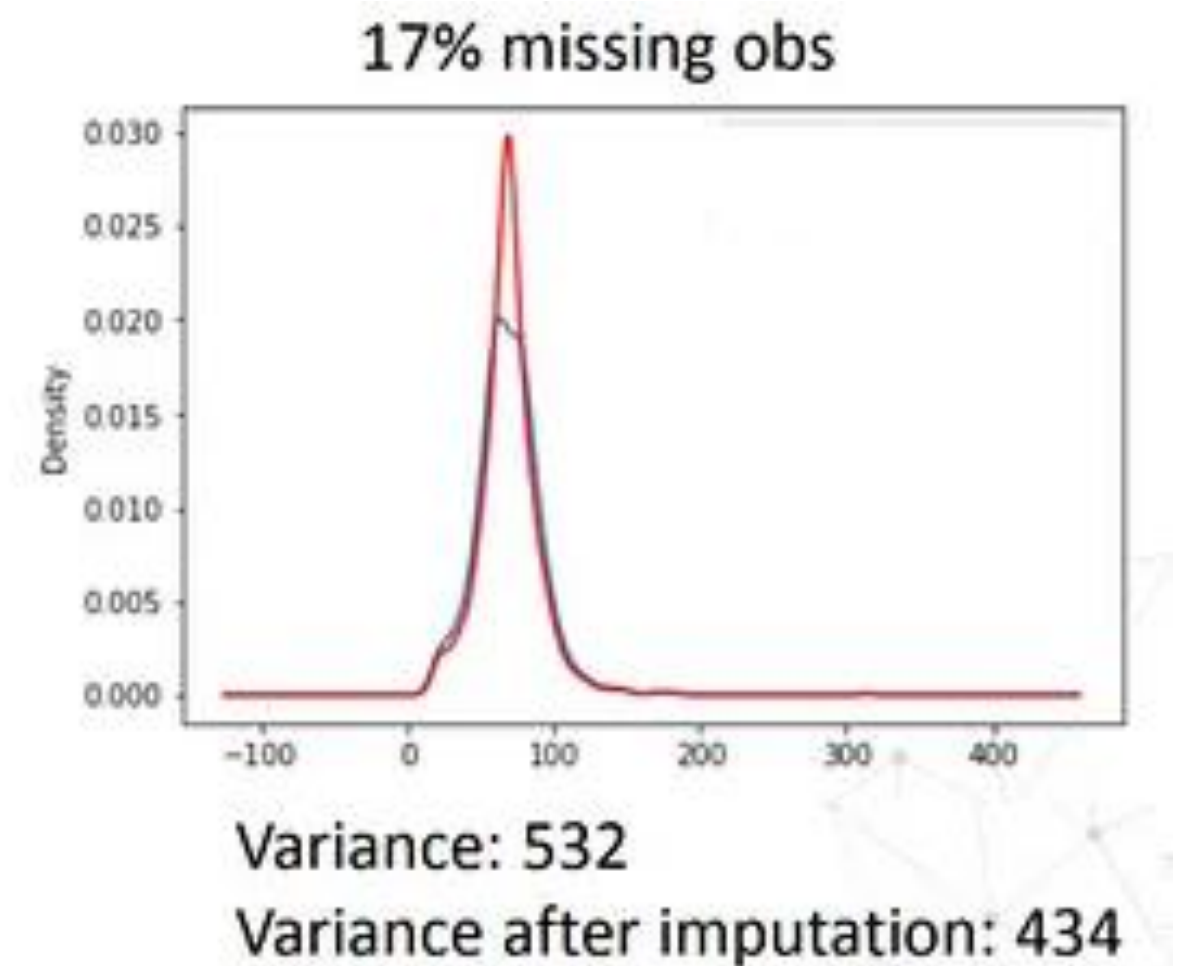
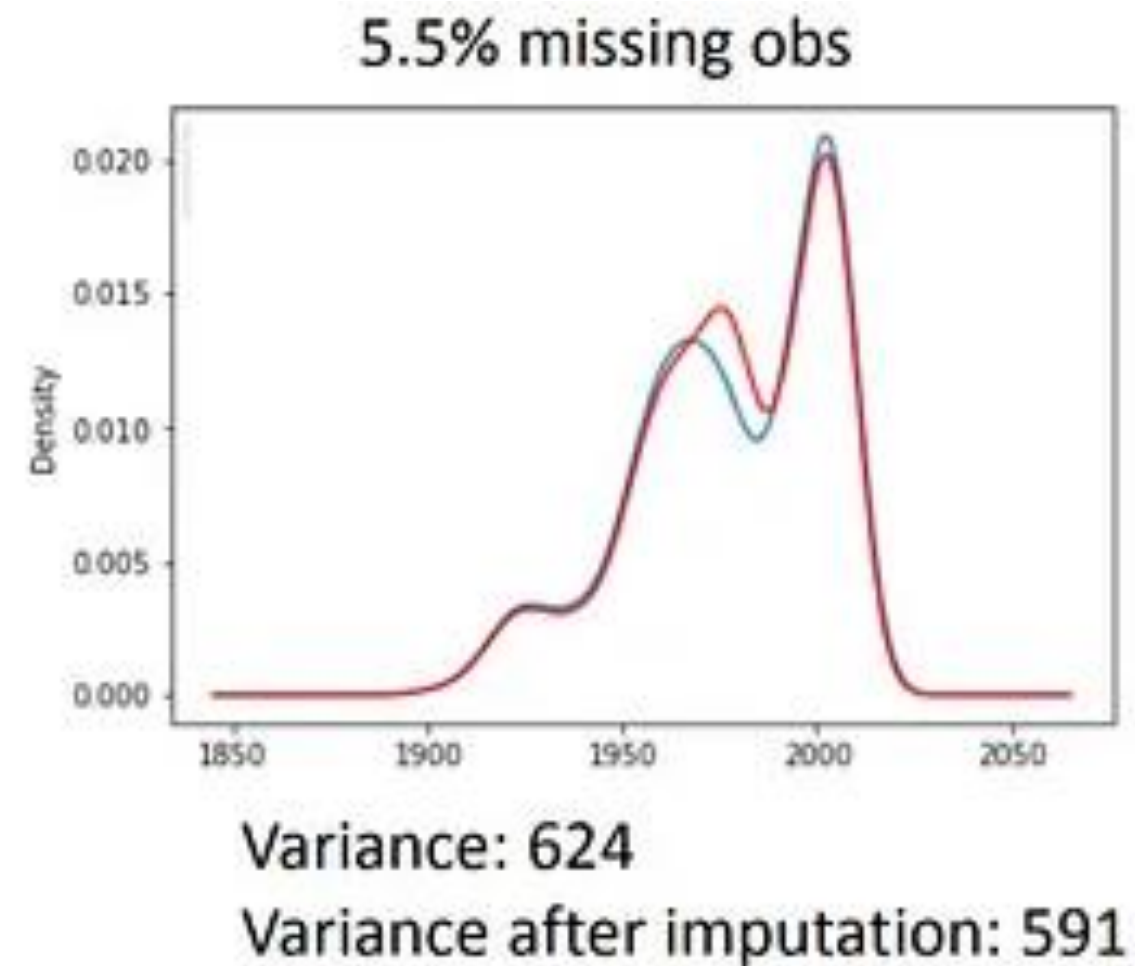
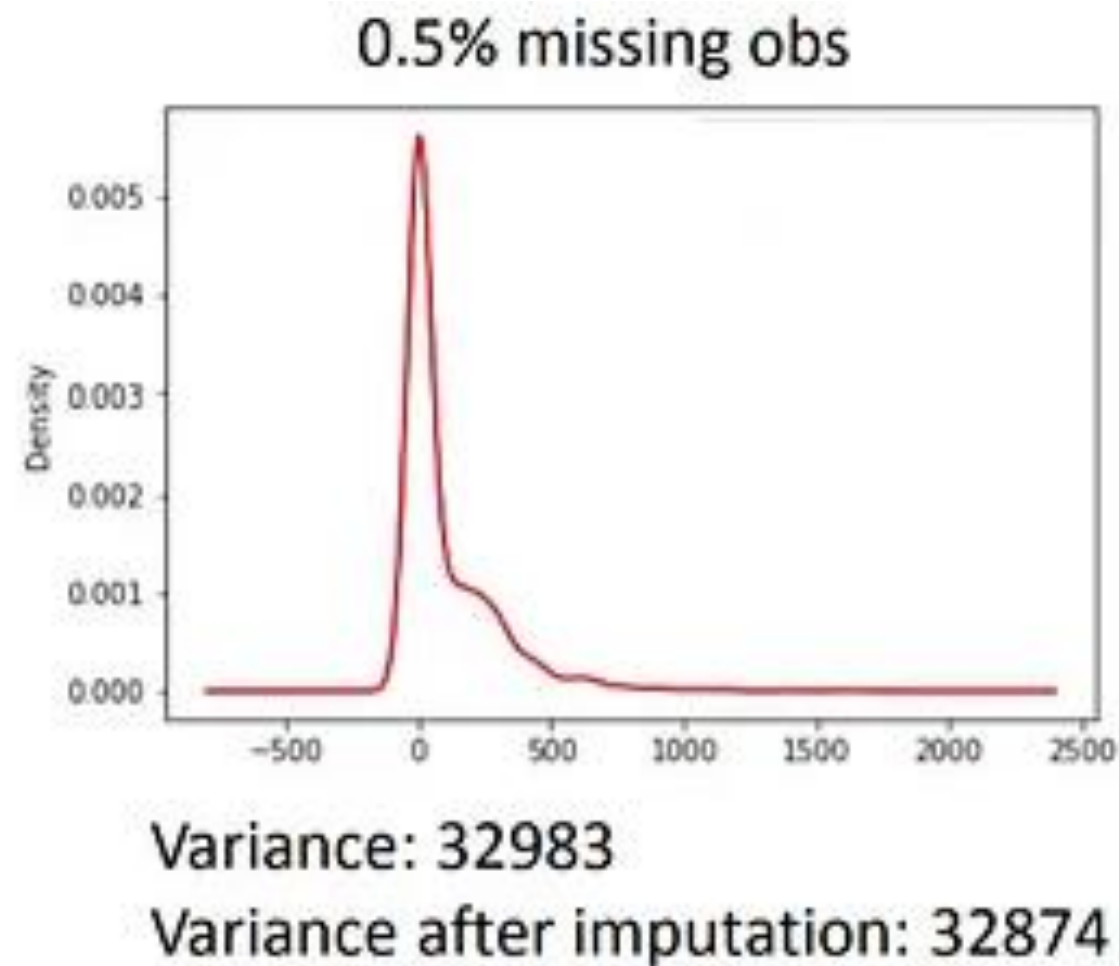
Median = 90



Price
100
90
50
40
20
100
86.66
60
120
86.66
200

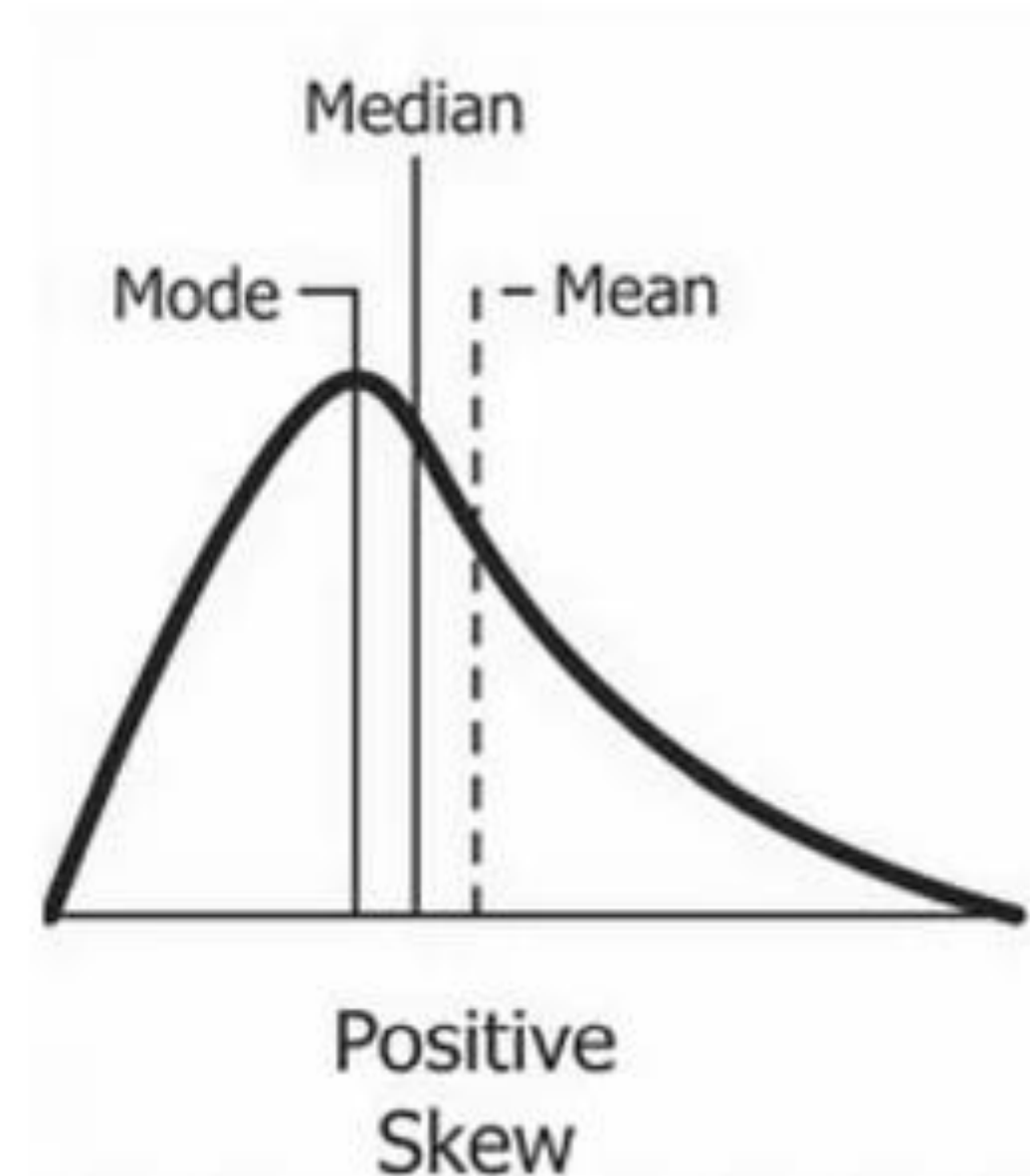
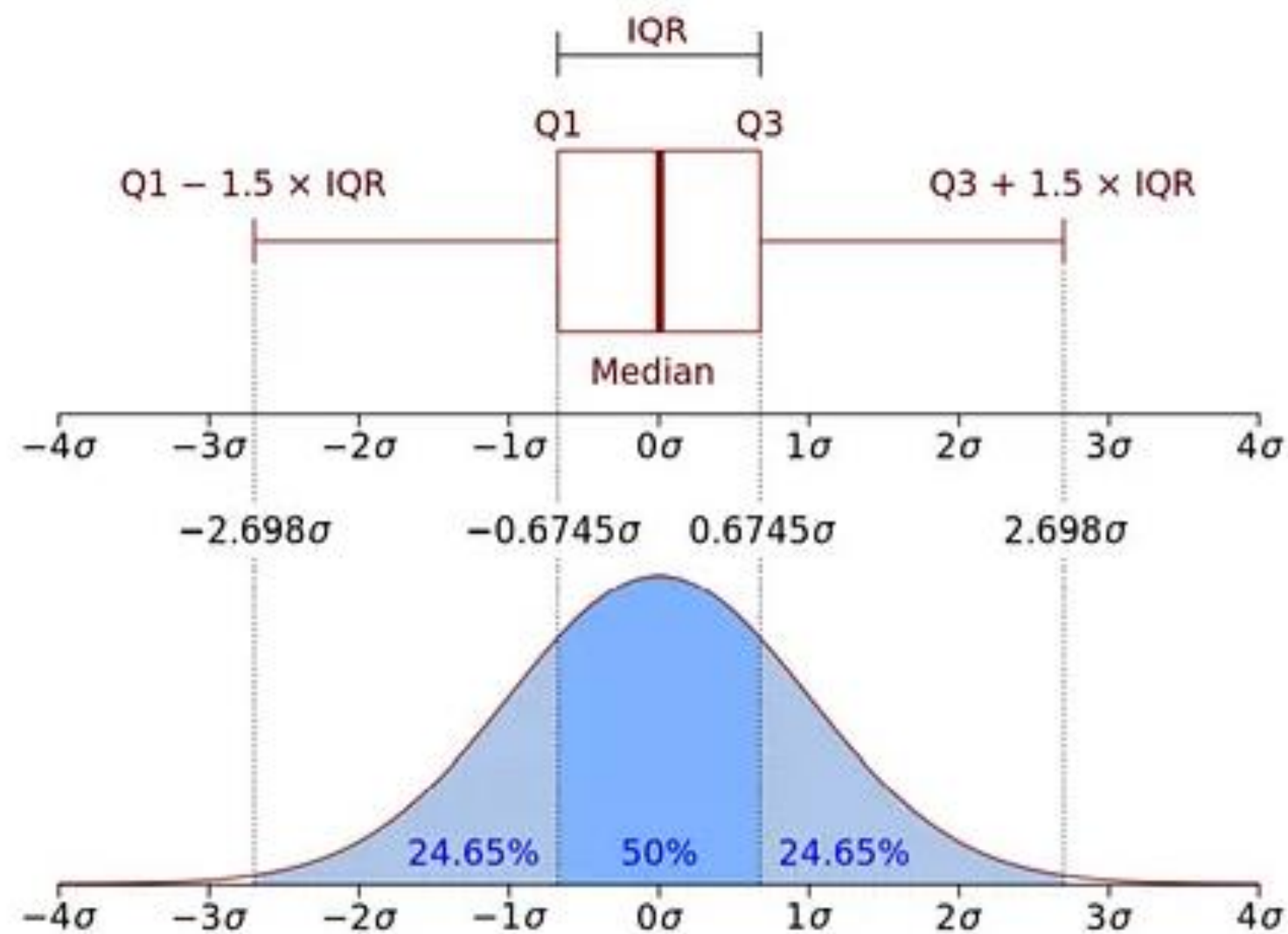
- Assumptions:
- Data is MCAR
- Missing data is mostly same as the rest

Impact of imputation on variance



- Variance decreases why?

Mean/Median Imputation When?



- If a feature is normally distributed, both mean & median imputation are equal
- If a feature is skewed or has many outliers, then median is better

Grouped Mean imputation

sepal length	sepal width	target
5.1	3.4	0
5.5	4.2	0
5.5	Nan	1
5.1	Nan	0
6.6	3.0	1
6.7	Nan	2
6.2	3.4	2

- Sepal width mean = 3.0
- Setosa Sepal width mean = 2.7
- Virginica Sepal width mean = 3.2
- Solution:
- Impute with grouped mean



Imputation using kNN

kNN for Imputation

HR	BP	Temp
76.0	126.0	38.0
74.0	120.0	NaN
72.0	118.0	37.5
NaN	136.0	37.0
77.0	NaN	39.0

- Logic: Nearest points share similar data
- kNN Imputation working summary:
- Choose $k=3$
- Find Euclidean distance between
 - Highlighted row & other rows
- Sort distance like kNN
- Pick the top 3 and average

Catch: Regular Euclidean distance wont work due to Nan

Nan Euclidean distance

- Distance between points (3, Nan, 5) & (1,0,0)

$$\sqrt{\frac{3}{2}\{(3-1)^2 + (5-0)^2\}} = 6.595453$$

$$d_{xy} = \sqrt{\text{weight} * \text{squared distance from present coordinates}}$$

$$\text{weight} = \frac{\text{Total number of coordinates}}{\text{Number of present coordinates}}$$

scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.nan_euclidean_distances.html

Distance between (100, Nan, 0.1) & (110, 0.3, 0.2)

All features should be on same scale
before KNN imputation

kNN Imputation with k=2

HR imputation for highlighted record with Nan adjusted distance

HR	BP	Temp
76.0	126.0	38.0
74.0	120.0	NaN
72.0	118.0	37.5
NaN	136.0	37.0
77.0	NaN	39.0

$$\sqrt{\frac{2}{2}} \times \{(136 - 126)^2 + (37 - 38)^2\}$$

$$\sqrt{\frac{2}{1}} \times \{(136 - 120)^2\}$$

$$\sqrt{\frac{2}{2}} \times \{(136 - 118)^2 + (37 - 37.5)^2\}$$

$$\sqrt{\frac{2}{1}} \times \{(37 - 39)^2\}$$

- Sort asc & pick top 2
- Average

kNN imputation considerations

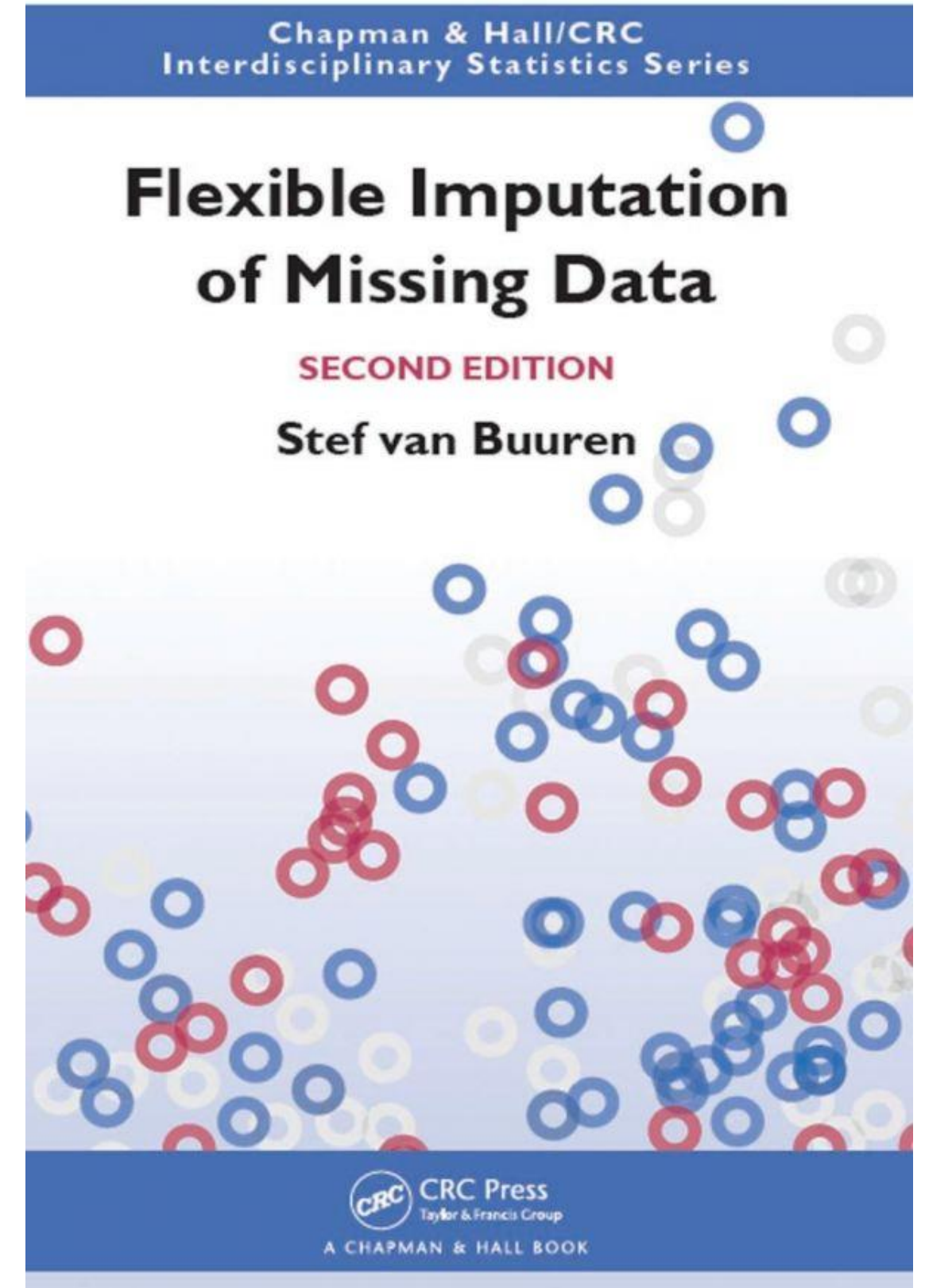
- If kNN is used for imputation, don't use kNN for classification/regression
 - E.g. Choose RandomForest Regressor for prediction
- Choosing k value for imputation is a part of cross validation

```
imputer = KNNImputer(n_neighbors=2, weights='uniform', metric='nan_euclidean')  
imputer.fit_transform(df)
```




Missing Data: The missing parts

- Imputation covered in future classes
 - Iterative Imputation,
 - MICE, MissForest
- Markov Chain Monte Carlo methods (will not be covered)





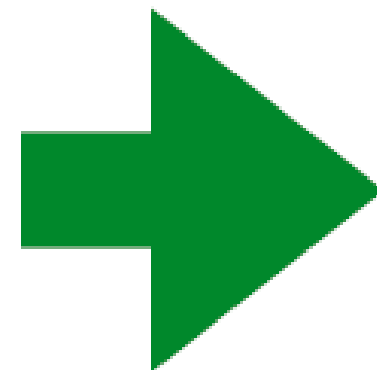
Categorical Variable encoding & imputation

	has heart disease?	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	no	55	no	nurse	pain	40s	133000
2	no	71	no	admin	beta blockers, pain	20s	34000
3	yes	89	yes	nurse	beta blockers	50s	40000
4	no	67	no	doctor	none	50s	120000

	has heart disease?	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)
1	no	55	no	nurse	pain	40s	133000
2	no	71	no	admin	beta blockers, pain	20s	34000
3	yes	89	yes	nurse	beta blockers	50s	40000
4	no	67	no	doctor	none	50s	120000

has heart disease?	
1	no
2	no
3	yes
4	no

$$\{\text{'yes'}, \text{'no'}\} \leftrightarrow \{+1, -1\}$$



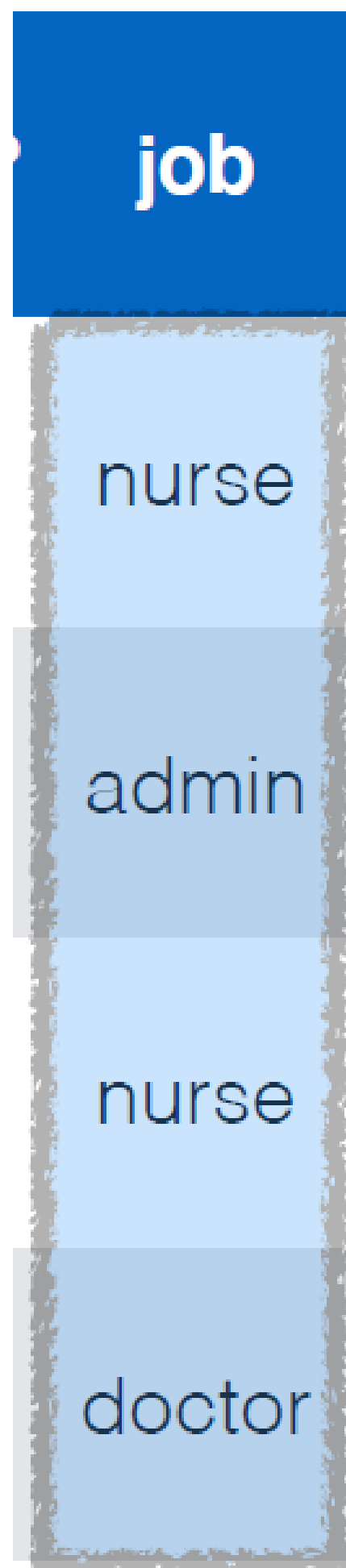
1	-1
2	-1
3	+1
4	-1

Can be mapped to 0 and 1

Depends on Algorithm

1. -1, +1 – Perceptron, SVM
2. 0, 1 Logistic Regression

	resting heart rate (bpm)	pain?	job	medicines	age	family income (USD)	pain?
1	55	no	nurse	pain	40s	133000	0
2	71	no	admin	beta blockers, pain	20s	34000	0
3	89	yes	nurse	beta blockers	50s	40000	1
4	67	no	doctor	none	50s	120000	0



- Ordinal Encoding - 1, 2, 3, 4
- Is admin > nurse and doctor
- Is it a linear scale?

- Binary code - 00, 01, 10, 11
- Inadvertently introduced pattern in job

nurse	0	0	0
admin	0	0	1
pharmacist	0	1	0
doctor	0	1	1
social worker	1	0	0

job
nurse
admin
nurse
doctor

- Turn each category into 0/1
- One hot encode – 0001 0010 0100 1000
- No pattern, feature explosion
- Not good for high cardinality

nurse	1	0	0	0	0
admin	0	1	0	0	0
pharmacist	0	0	1	0	0
doctor	0	0	0	1	0
social worker	0	0	0	0	1

medicines

pain

beta blockers,
pain

beta blockers

none

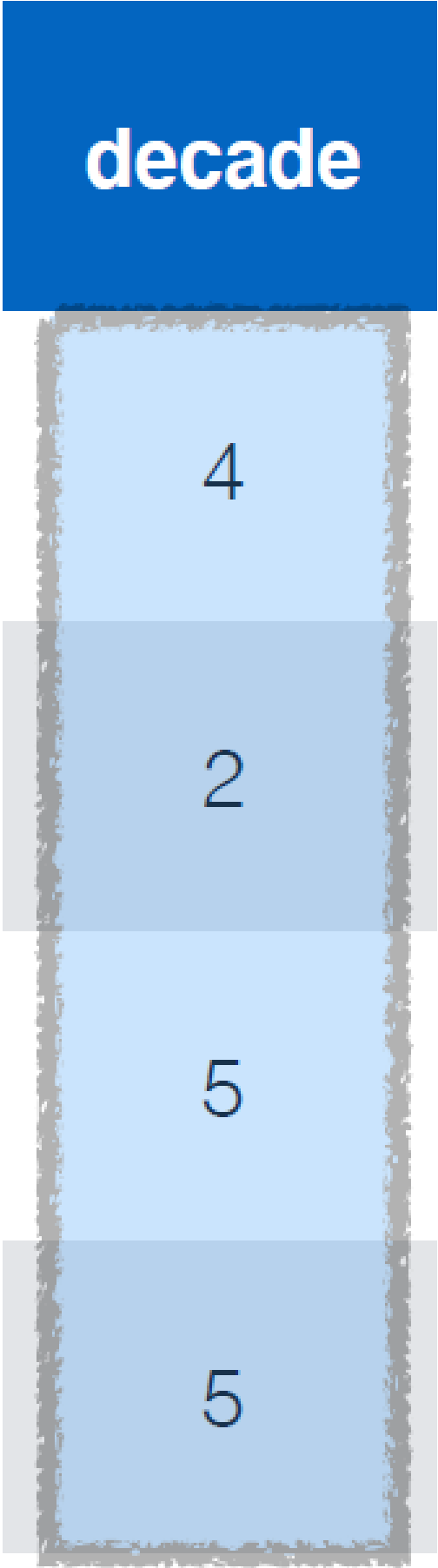
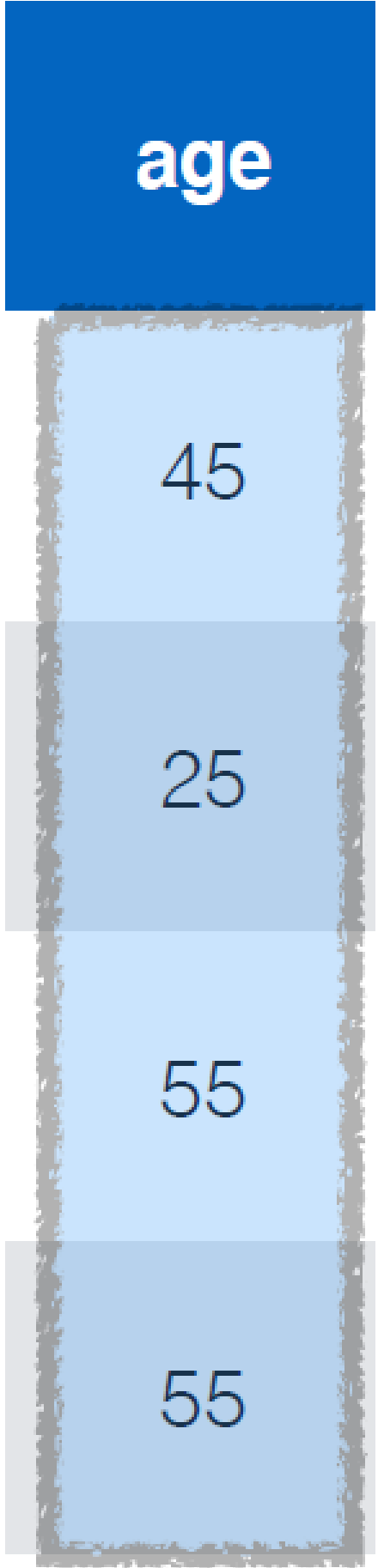
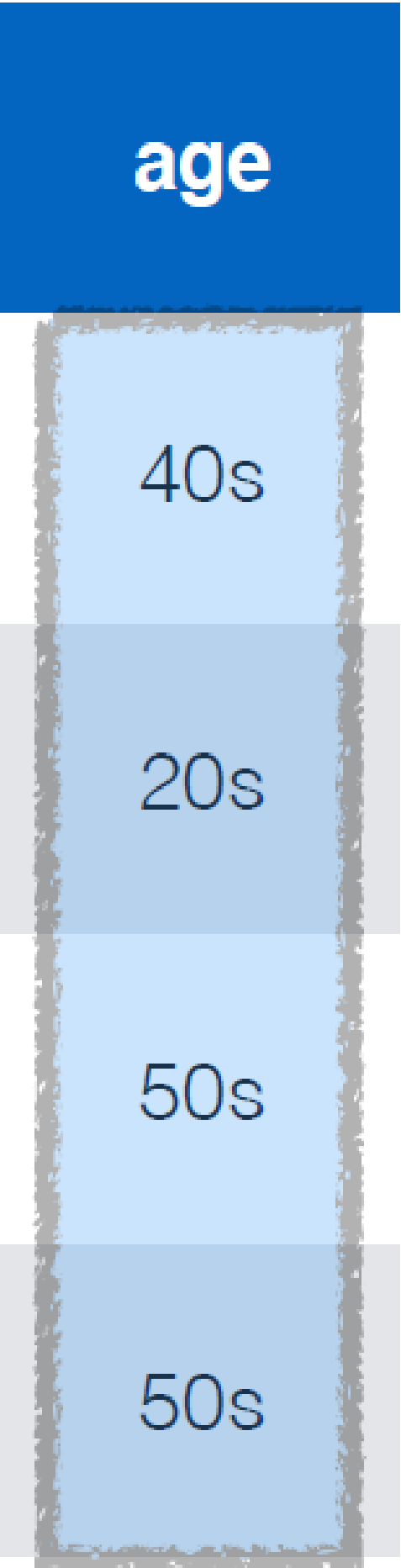
- Should we use one hot encoding?

pain	1	0	0	0
pain & beta blockers	0	1	0	0
beta blockers	0	0	1	0
no medications	0	0	0	1

- How about Factored encoding?

pain	1	0
pain & beta blockers	1	1
beta blockers	0	1
no medications	0	0

- How is it different from binary encoding?



**Reminds of IP
Address Geo
location fiasco**

Encode Ordinal Data

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	2	3	4	5

- Log Levels
 - Trace, Debug, Info, Warn, Error, Fatal
 - Do these map to 1,2,3,4,5,6?
 - Or 1,2,5,8,10?

Imputing Categorical Variables

- Cannot do mean/median
- Impute the most frequent value
- Grouped mode
- Impute the value to be a new category

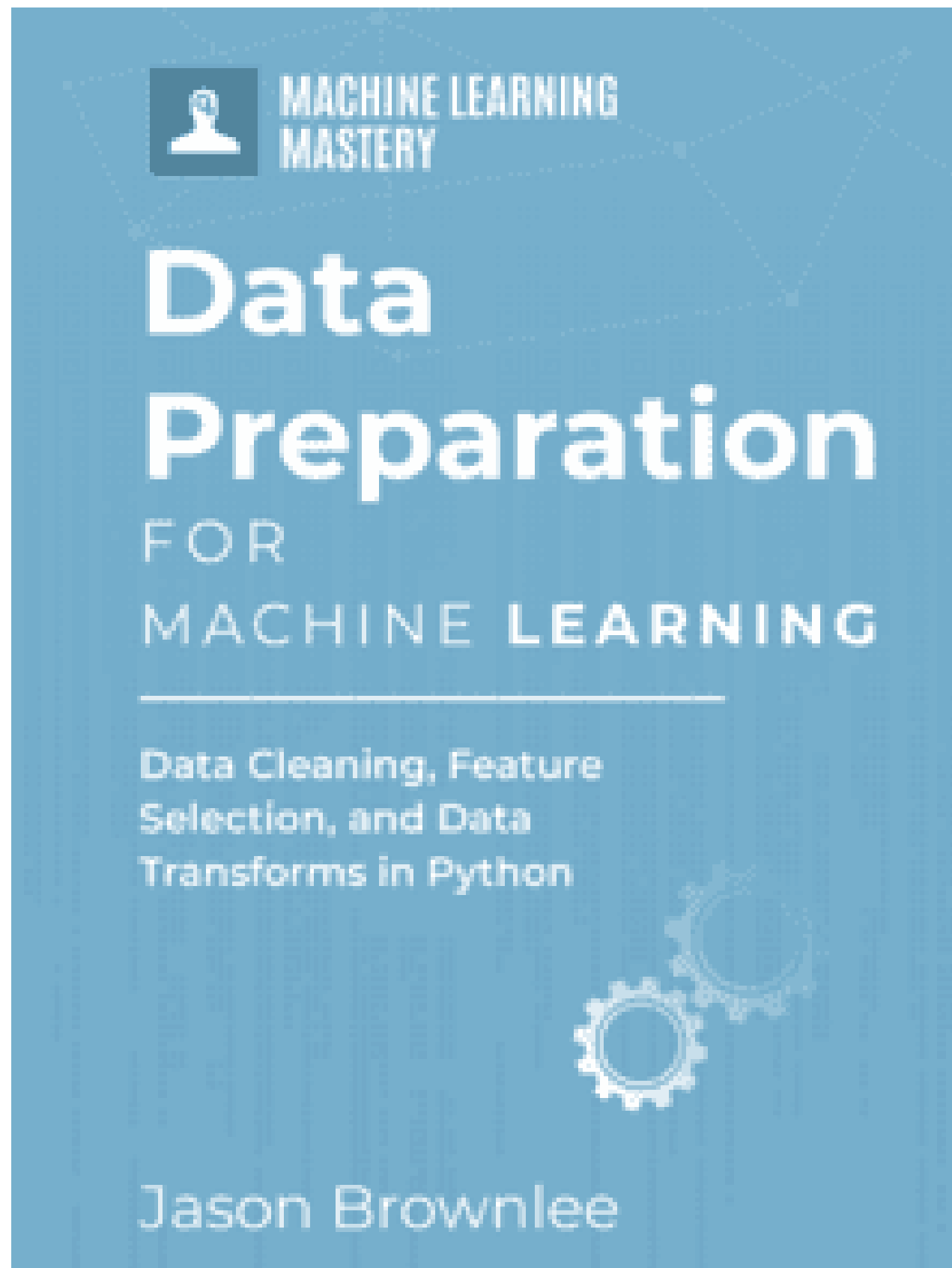
Recap

- kNN Regression
- Data imputation
 - Univariate – Mean, Median, Mode, Grouped Mean
 - Multivariate – kNN (Nan Euclidean distance)
- Encoding transformations for categorical variables
 - Ordinal, Factored, One hot, binary,
- Problem: Imputation first or standardization first?

Solutions: No one right answer

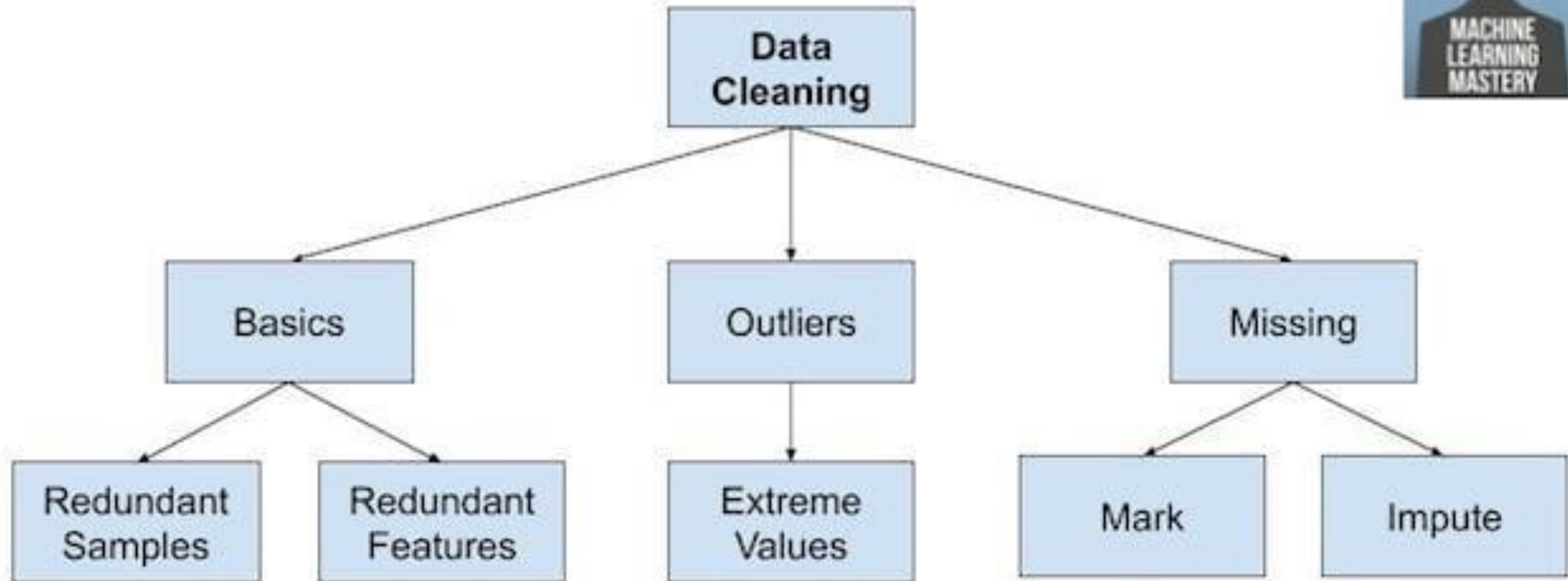
- Some have opined standardize first
- Others: impute with a temporary value first
- Hybrid solution approach (one variation):
 - Grouped mean impute + missing indicator (adds 0/1)
 - Standardize
 - Set Nan again using Missing Indicator, then KNN Impute
- Can use missing indicator as feature for ML prediction

Data cleaning, Feature Engineering books



Data cleaning: Big picture

Overview of Data Cleaning





QUESTIONS



Thank You!