

# Data wrangling and preprocessing



I HEARD YOU HAVE SOME  
DIRTY DATA TO CLEANSE.

# The world is a messy place!

**Just because you have data, doesn't make it useful.**

- Real-world data is not clean.
- Some data points might be missing.
- Some others might be out of range.
- There could be duplicates.
- The data might come from different sources in different formats.



Dataedo /cartoon

Piotr@Dataedo

**Garbage in, garbage out: Low quality data leads to flawed analysis**

# Data Wrangling

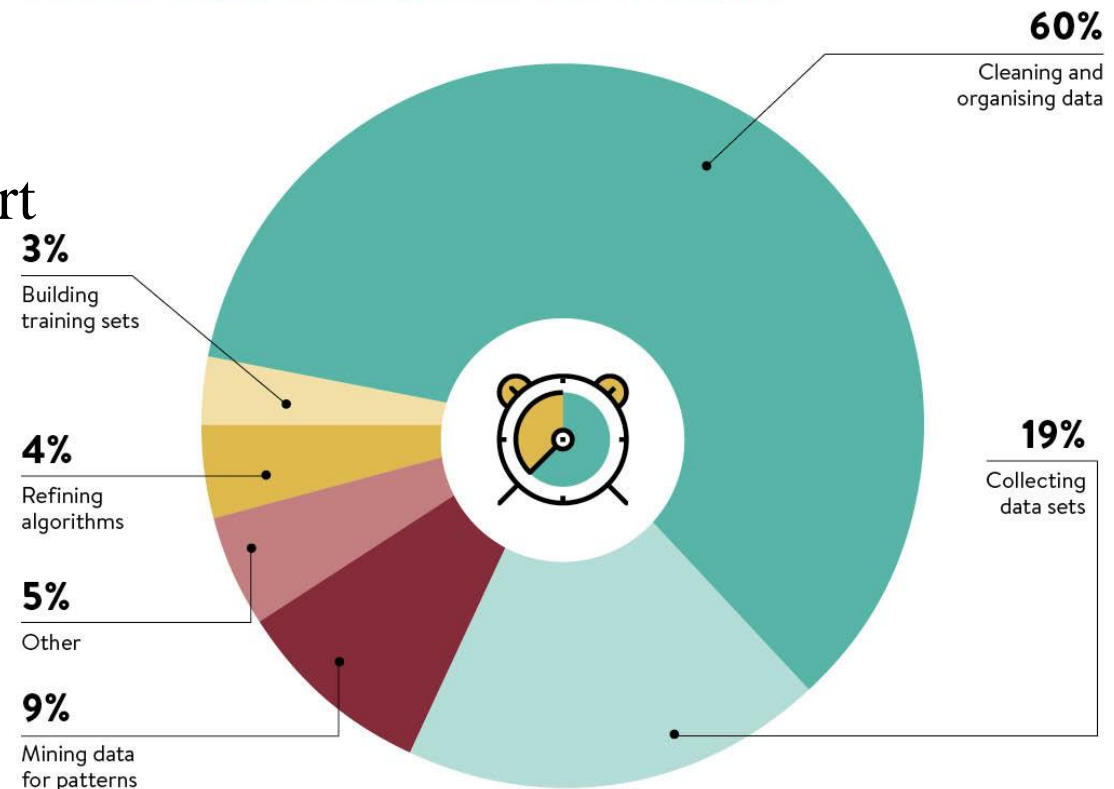
- **Data wrangling**— also called *data cleaning*, *data remediation*, or *data munging*— refers to a variety of processes designed to transform raw data into more readily used formats.
- Data wrangling is about **gathering** the right pieces of data, **assessing** your data's quality and structure, then modifying your data to make it **clean**.
- An iterative process between:
  - Gathering;
  - Assessing, and;
  - Cleaning Data

# The 80-20 rule of data wrangling

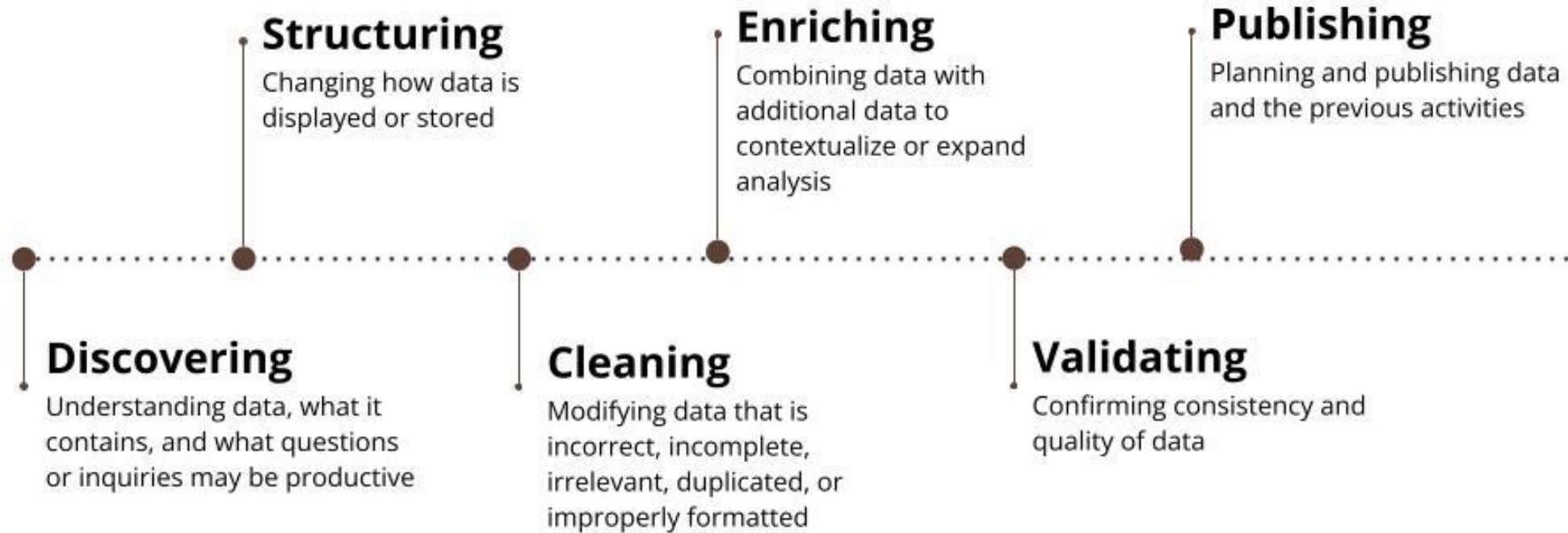
80% of a data scientist's time and effort is spent in collecting, cleaning and preparing the data

76% of data scientists view data preparation as the least enjoyable part of their work.

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



# The Data Wrangling Process



# Defining terms

- A **dataset** is collection of **data objects**.
- A data object represents an entity
  - Can also be referred to as *samples, examples, instances, data points, or objects*.
  - Examples:
    - customers, store items, and sales in a sales database
    - patients in a medical database
    - students, professors, and courses in a university database
- **Data objects** are typically described by **attributes**

# Defining terms

- Majority of Data Mining works assumes that data is a collection of records.
  - Other types of datasets exist (graph-based, ordered, spatial..)
- Record data has no explicit relationship among records or data fields, and every record has the same set of attributes.

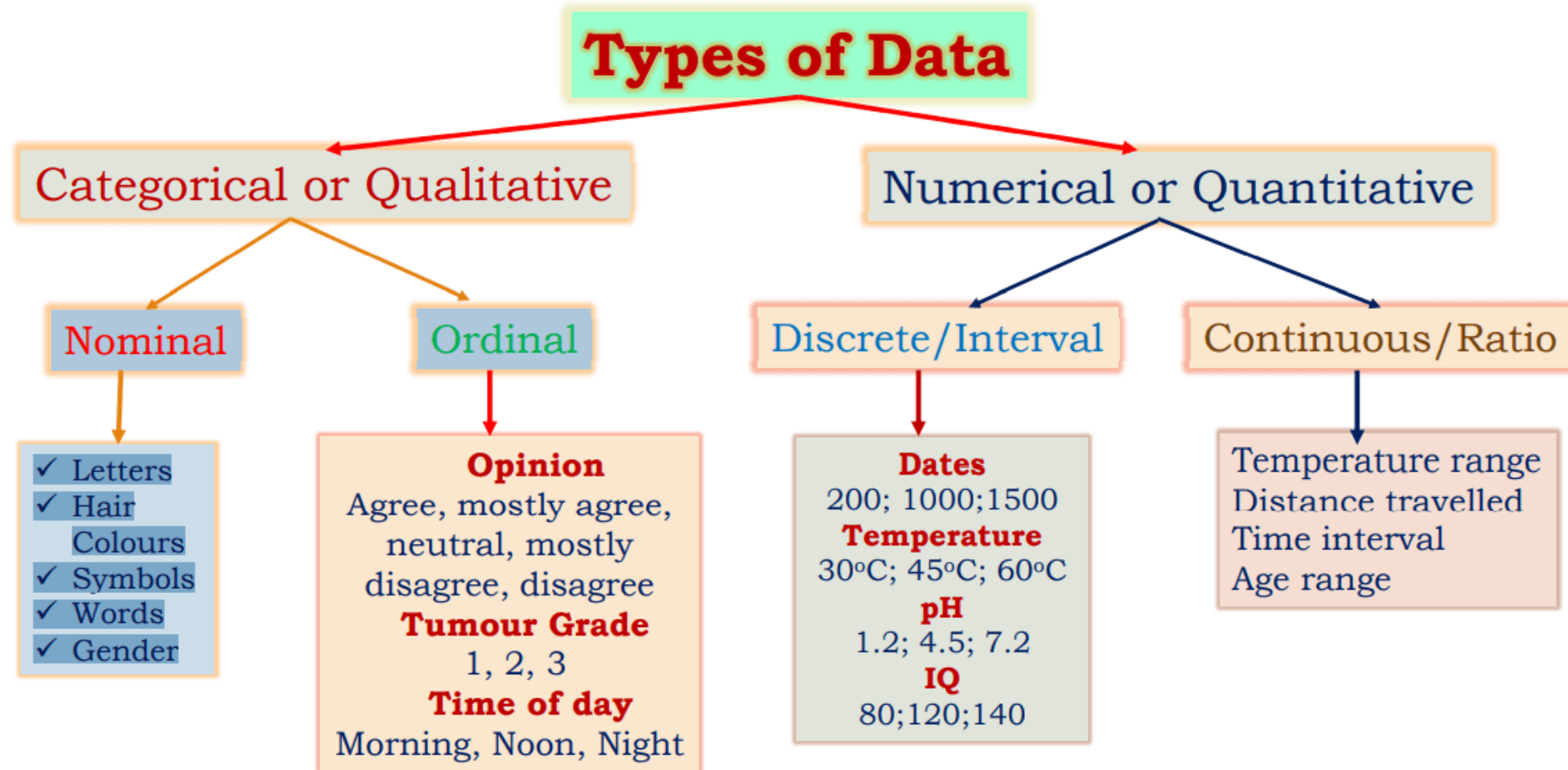
	A	B	C	D	E	F
1	First Name	Last Name	Date of birth	Age	Salary	Department
2	Hank	McNeil	1-2-1993	25	€ 20.000,00	Sales
3	Jessica	Williams	15-4-1956	62	€ 35.000,00	R&D
4	Rick	Johnson	30-6-1966	52	€ 40.000,00	Management
5	John	Jenkins	17-4-1969	49	€ 30.000,00	Sales
6	Joe	Vanderberg	4-11-1970	48	€ 32.000,00	Sales
7	Mary	Dylan	12-12-1979	39	€ 60.000,00	Management
8	Leeroy	Johanson	12-7-1984	34	€ 24.000,00	R&D
-						

# Defining terms

- A data **attribute** is a data field, representing a characteristic or feature of a data object.
  - The nouns *attribute*, *dimension*, *feature*, and *variable* are often used interchangeably
- The type of an attribute is determined by the set of possible values—nominal, binary, ordinal, or numeric—the attribute can have.
  - Statistical methods (used in data exploration) can only be used with certain data types.
  - Knowing the types of data you are dealing with, enables you to choose the correct method of analysis.



# Defining terms



# Defining terms

There are two basic data types, each with two sub-types

□ **Numerical**: expressed by numbers

- Discrete: numbers take on integer values only (e.g. # of children, # of siblings)
- Continuous: numbers can take on decimal values (e.g. height, weight)

□ **Categorical**: expressed by categories (also known as factors/groups)

- Nominal: no meaningful order between categories (e.g. gender, occupation)
- Ordinal: categories can be put in meaningful order (e.g. agreement, level of pain, etc.)

# Defining terms

<i>Amount of money earned last week</i> <i>Arm span</i> <i>Birthdate</i> <i>Concentration exercise (seconds)</i> <i>Dominant hand reaction time</i> <i>Favourite sport</i> <i>Height</i> <i>Hours slept per night</i>	<i>Language mostly spoken at home</i> <i>Foot length</i> <i>Opinions on environmental conservation</i> <i>School post code</i> <i>State/Territory live in</i> <i>Travel method to school</i> <i>Travel time to school</i> <i>Year level</i>
<u>Categorical</u>	<u>Numerical</u>

# Defining terms

<i>Amount of money earned last week</i> <i>Arm span</i> <i>Birthdate</i> <i>Concentration exercise (seconds)</i> <i>Dominant hand reaction time</i> <i>Favourite sport</i> <i>Height</i> <i>Hours slept per night</i>	<i>Language mostly spoken at home</i> <i>Foot length</i> <i>Opinions on environmental conservation</i> <i>School post code</i> <i>State/Territory live in</i> <i>Travel method to school</i> <i>Travel time to school</i> <i>Year level</i>
<p style="text-align: center;"><b><u>Categorical</u></b></p> <ol style="list-style-type: none"> <li>1. Birthdate</li> <li>2. Favourite sport</li> <li>3. Language mostly spoken at home</li> <li>4. Opinions on environmental conservation</li> <li>5. School post code</li> <li>6. State/Territory live in</li> <li>7. Travel method to school</li> <li>8. Year level</li> </ol>	<p style="text-align: center;"><b><u>Numerical</u></b></p> <ol style="list-style-type: none"> <li>1. Amount of money earned last week</li> <li>2. Arm span</li> <li>3. Concentration exercise (seconds)</li> <li>4. Dominant hand reaction time</li> <li>5. Height</li> <li>6. Hours slept per night</li> <li>7. Foot length</li> <li>8. Travel time to school</li> </ol>

# First look at your data

- Get familiar with the data. Take a look at the data you have and think about how you would like it organized to make it easier to analyze.
- Identify and detect obvious issues or mistakes that need to be addressed.
  - Missing or incomplete values
  - Extreme or outliers values
  - Invalid or unusual values
  - Non standard units (km, meters, inches, etc. all mixed)
  - Innacurate data, inconsistent data, etc.

# Data issues

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL
...						

# Data issues

## Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL
...						

# Data issues

## Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL
...						

Missing Values



# Data issues

## Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL
...						

Typos

Missing Values

# Data issues

## Duplicates

## Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL
...						

## Typos

## Missing Values

# Data issues

## Duplicates

## Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

Maybe Duplicates?

Typos

Missing Values

# Structuring data

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

—HADLEY WICKHAM

## In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable



id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation



# Data cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy

# Missing data

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

Name	Age	Sex	Income	Class
Mike	40	Male	150k	Big spender
Jenny	20	Female	?	Regular
...				

Customer Data

# Types of Missing data

- **Type 1: Missing Completely At Random (MCAR)**
  - The missing data are unrelated to the observation being studied or the other variables in the data set.
  - The data of interest is not systematically different between missing and observed.

MCAR		
ID	Gender	Depression Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8
Missing Value		

# Types of Missing data

- **Type 2: Missing At Random (MAR)**
  - The fact that data are missing can be predicted from the other variables in the study, but not from the missing data themselves.
  - Whether or not a data point is missing is dependent on observed data.

MAR		
ID	Gender	Depression Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8
Missing Value		



# Types of Missing data

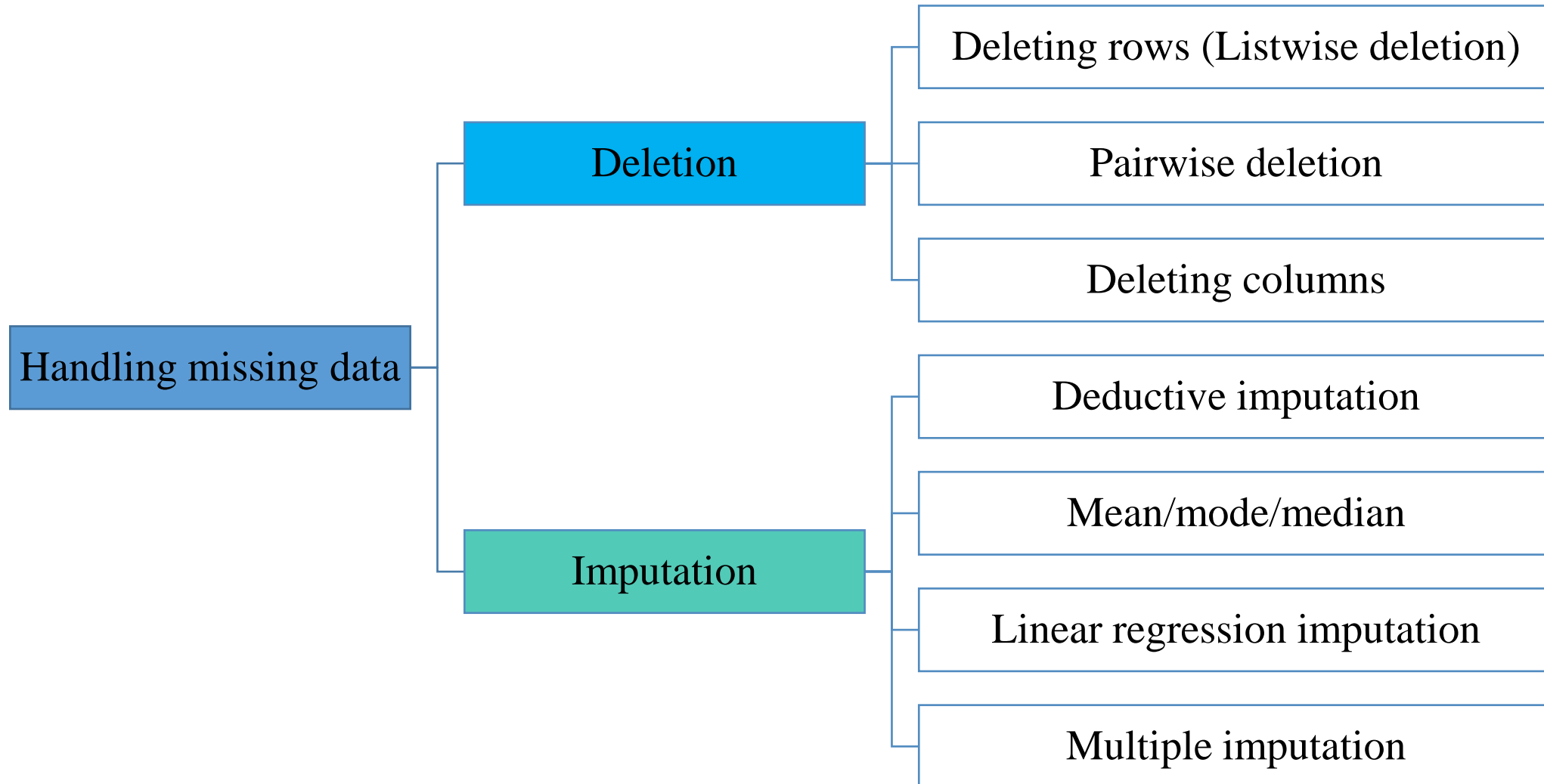
- **Type 3: Missing Not At Random (MNAR)**
  - The data of interest are systematically different for missing and observed.
  - Whether or not an observation is missing depends on the value of the unobserved data itself

MNAR		
ID	Gender	Depression Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8
Missing Value		

# How to Handle Missing Data?

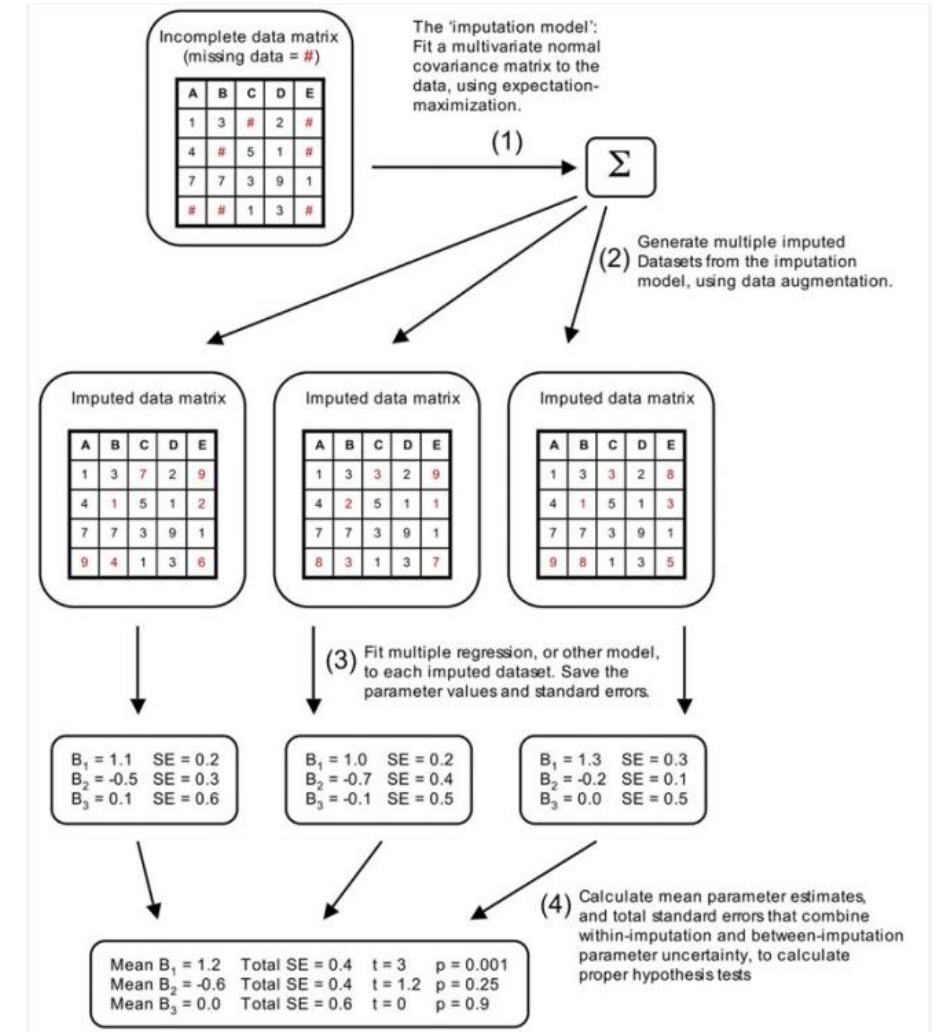
- **Ignore:** usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually:** tedious + infeasible?
- **Fill in it automatically** with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: *smarter*
  - the most probable value: inference-based such as Bayesian formula or decision tree

# Handling Missing data



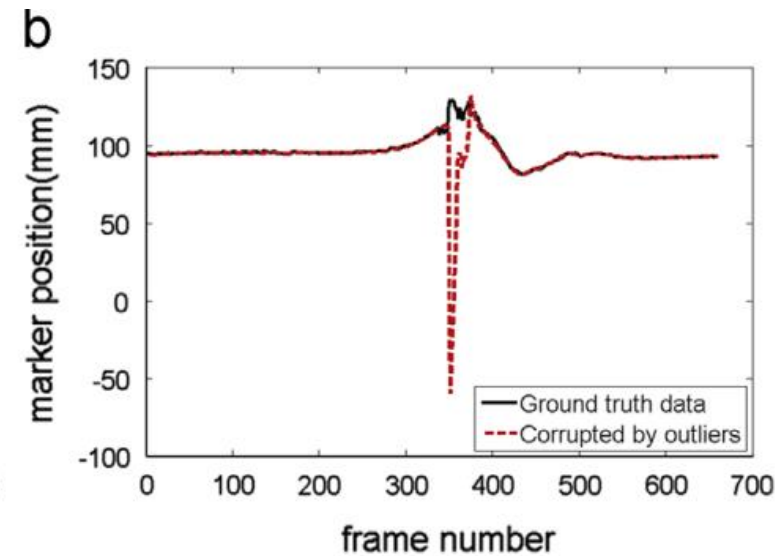
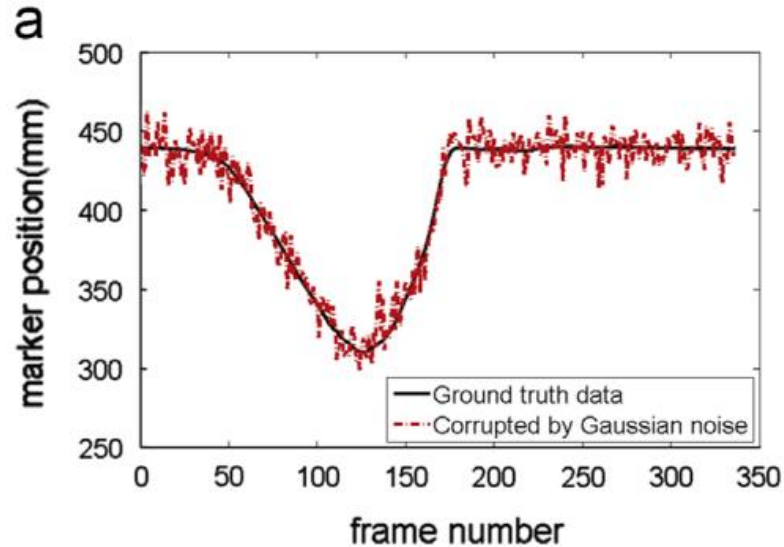
# Handling Missing data

- Multiple imputation
  - Make multiple copies of the dataset.
  - Use imputation to generate one value for each NA in each dataset.
  - Once imputation of all datasets is complete, do a “final model” or “final analysis” on each dataset.
  - then combine the results of the multiple models together, just like we aggregate results in an ensemble model



# Noise/outlier data

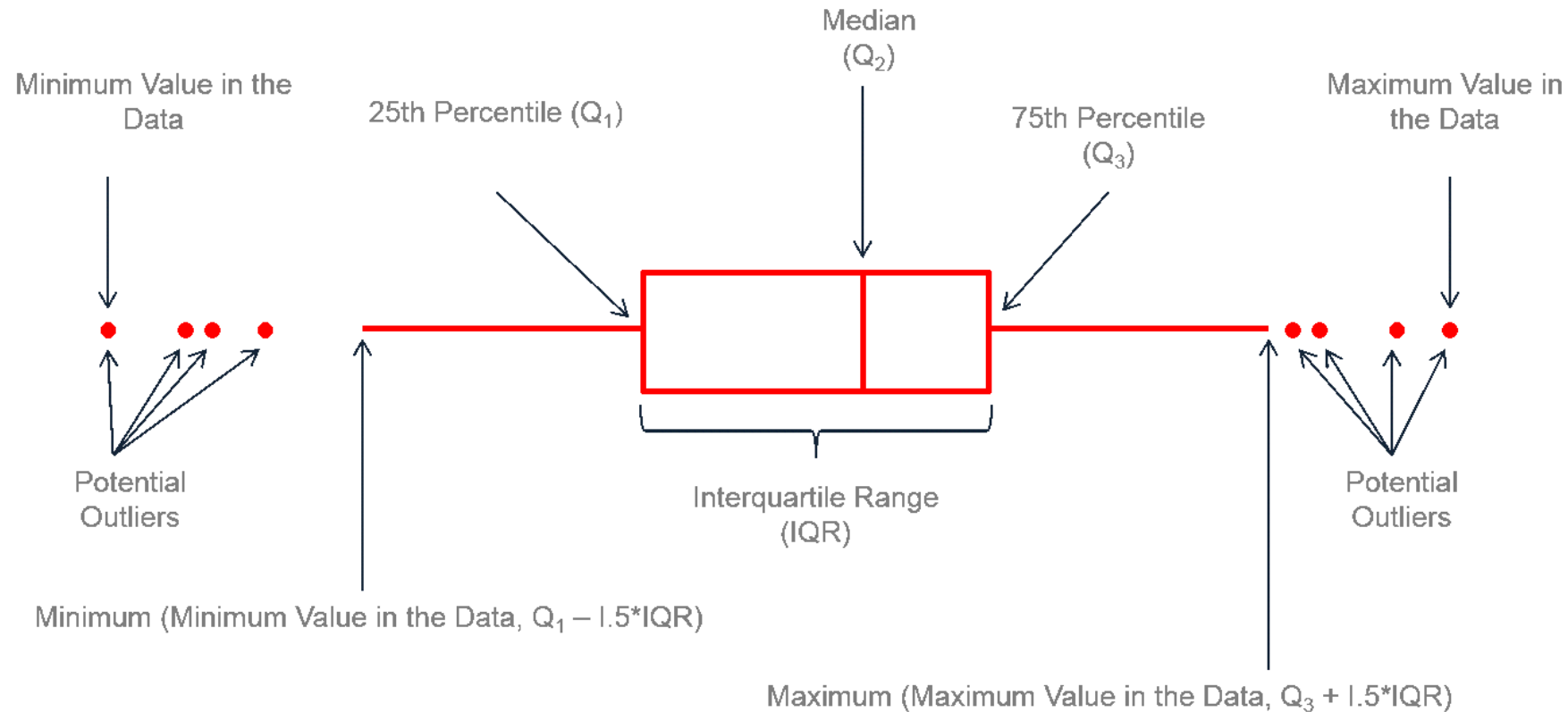
- Noise and outliers are ‘bad measurements’ in data
  - Random errors or variance in a measured variable
- Noise and outliers can negatively influence your data analysis, so it is better to have them filtered out



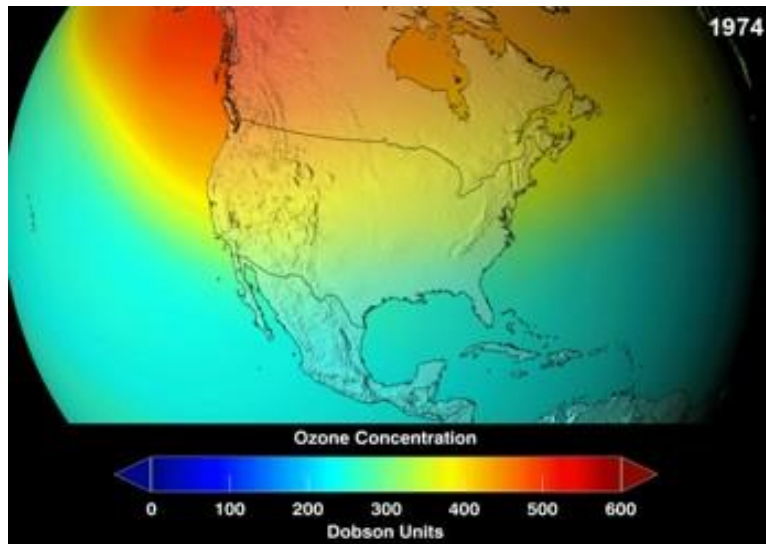
- **Outliers could be valid but rare data**

# Outlier identification

- Potential outlier identification



# Sometimes outliers are important data!



Always always  
always! Look at  
the data!

The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin...came as a shock to the scientific community...[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

Ozone depletion#Antarctic ozone hole

# Handling noisy data

## ❑ Binning

- smoothing a sorted data value by consulting its “neighborhood,” that is, the values around it.
- Sorted is partitioned into bins (equal-frequency or equal-depth), then smoothed by bin means, bin median, bin boundaries, etc.

## ❑ Regression

- smoothing by fitting the data into a regression function.

## ❑ Clustering:

- similar values are organized into groups (clusters) and values that fall outside of the set of clusters may be considered outliers



# Handling outliers

## ❑ **Removal:**

- Discarding should be approached with caution, but sometimes there are valid reasons to remove outliers (E.g. measurement errors (a pulse of 0, or a “teen” with an age of 155))

## ❑ **Transformation:**

- Use scaling or normalization methods like min-max scaling, standardization, or log transformation.

## ❑ **Binning:**

- Group outliers into a separate category or bin to treat them separately during analysis.

## ❑ **Imputation:**

- Replace outliers with more reasonable values, such as the median or mean of the dataset.

# Data transformation (Feature transformation)

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
- Among the techniques :
  - Feature scaling
  - Feature encoding (for categorical features)
  - Log transform

# Feature scaling

- **Min-max normalization** to  $[\text{new\_min}_A, \text{new\_max}_A]$  (also known as **normalization**)

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0].

Then \$73,000 is mapped to  $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$

- **Z-score normalization** ( $\mu$ : mean,  $\sigma$ : standard deviation): (also know as **standardization**)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Ex. Let  $\mu = 54,000$ ,  $\sigma = 16,000$ . Then  $\frac{73,600 - 54,000}{16,000} = 1.225$

# Feature scaling : normalize or standardize?

- Normalization is good when :
  - you know that the distribution of your data **does not follow a Gaussian distribution**. This can be useful in algorithms that do not assume any distribution of the data, like K-Nearest Neighbors and Neural Networks.
- Standardization can be helpful in cases where :
  - the data **follows a Gaussian distribution**. However, this does not have to be necessarily true.
  - Outliers will not be affected by standardization since that unlike normalization, standardization does not have a bounding range .

**No hard or fast rule!! always start by fitting your model to raw, normalized, and standardized data and comparing the performance for the best results.**

# Feature encoding

- We cannot feed categorical features as string types directly into models. Encoding methods are used to convert these features numerical.
- **Label/ordinal encoding:** assigns each categorical value an integer value based on alphabetical order.

Original Data		Label Encoded Data	
Team	Points	Team	Points
A	25	0	25
A	12	0	12
B	15	1	15
B	14	1	14
B	19	1	19
B	23	1	23
C	25	2	25
C	29	2	29

# Feature encoding

- **One hot encoding:** Transform a categorical feature with  $m$  possible values into  $m$  binary features

Customer	Car	----->	Customer	Bmw	Mercedes	Audi
1	Bmw		1	1	0	0
2	Mercedes		2	0	1	0
3	Bmw		3	1	0	0
4	Audi		4	0	0	1
5	Bmw		5	1	0	0

# Feature encoding

- **Frequency / Count Encoding** : Replace the categories by the count of the observations that show that category in the dataset.

	player	point
0	Stephen Curry	4
1	Anthony Edwards	6
2	Anthony Edwards	1
3	Duncan Robinson	2
4	Duncan Robinson	3
5	Stephen Curry	5
6	Stephen Curry	2
7	Stephen Curry	3



	player	point
0	4	4
1	2	6
2	2	1
3	2	2
4	2	3
5	4	5
6	4	2
7	4	3

# Feature encoding

- **Feature hashing:** Encoding categorical variables with the help of a hash function.
  - Very useful for large categorical variables. Represents data in a high-dimensional space using a fixed-size array.

	player	point
0	Stephen Curry	4
1	Anthony Edwards	6
2	Anthony Edwards	1
3	Duncan Robinson	2
4	Duncan Robinson	3
5	Stephen Curry	5
6	Stephen Curry	2
7	Stephen Curry	3

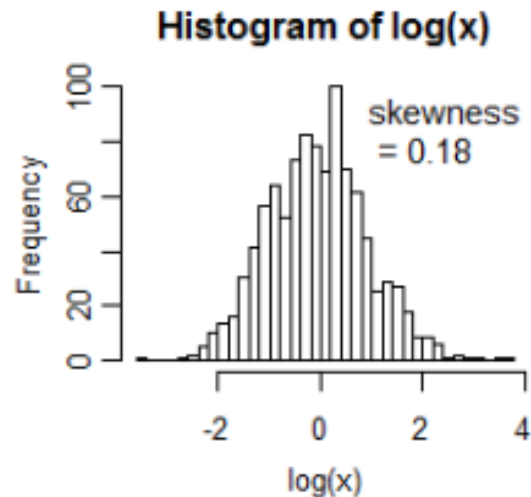
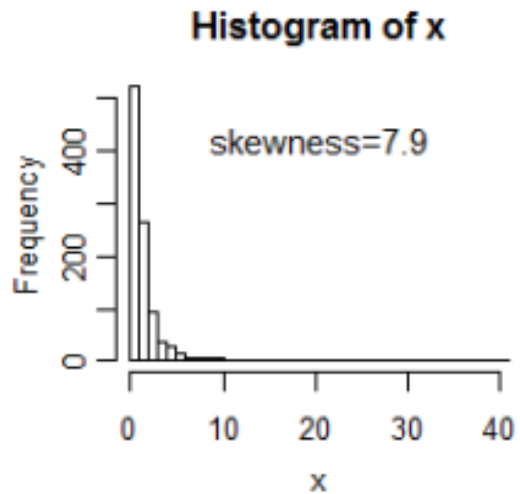


	col_0	col_1	col_2	col_3	col_4	col_5	col_6	col_7	point
0	0	0	0	0	1	0	0	0	4
1	0	0	1	0	0	0	0	0	6
2	0	0	1	0	0	0	0	0	1
3	0	0	1	0	0	0	0	0	2
4	0	0	1	0	0	0	0	0	3
5	0	0	0	0	1	0	0	0	5
6	0	0	0	0	1	0	0	0	2
7	0	0	0	0	1	0	0	0	3



# Log transformation

- Compresses the range of large numbers and expand the range of small numbers.
- It's mostly used to turn a skewed distribution into a normal or less-skewed distribution.

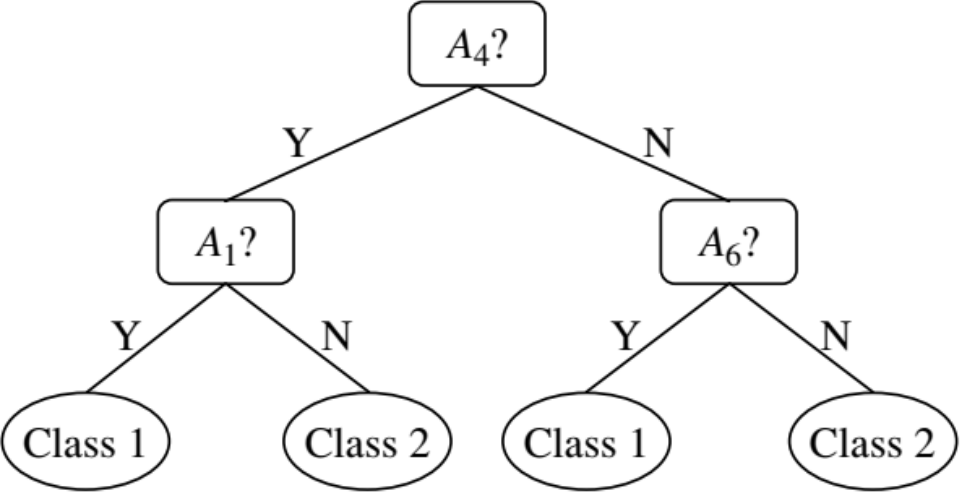


Original Data		Transformed Data	
Age X	Income Y	Age X	Income ln(Y)
25	18,500	25	$\ln(18,500)=9.83$
30	23,600	30	$\ln(23,600)=10.07$
35	29,800	35	$\ln(29,800)=10.30$
40	38,500	40	$\ln(38,500)=10.56$
45	49,000	45	$\ln(49,000)=10.80$
50	64,100	50	$\ln(64,100)=11.07$
55	78,500	55	$\ln(78,500)=11.27$
60	102,000	60	$\ln(102,000)=11.53$
65	130,800	65	$\ln(130,800)=11.78$

# Feature selection

- Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
- Reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.
- “How can we find a ‘good’ subset of the original attributes?”
  - Forward selection
  - Backward elimination
  - Decision tree induction

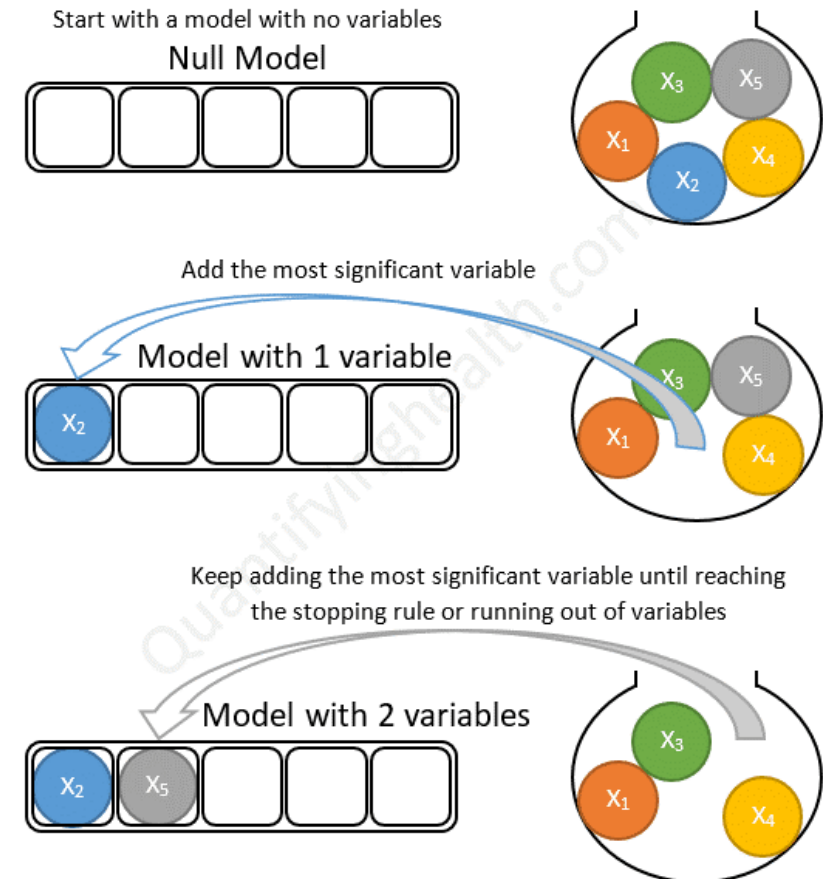
# Feature selection

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p>Initial reduced set:  <math>\{\}</math>  <math>\Rightarrow \{A_1\}</math>  <math>\Rightarrow \{A_1, A_4\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p><math>\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}</math>  <math>\Rightarrow \{A_1, A_4, A_5, A_6\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p>  <pre> graph TD     A4["A4?"] -- Y --&gt; A1["A1?"]     A4 -- N --&gt; A6["A6?"]     A1 -- Y --&gt; C1_1((Class 1))     A1 -- N --&gt; C2_1((Class 2))     A6 -- Y --&gt; C1_2((Class 1))     A6 -- N --&gt; C2_2((Class 2))     </pre> <p><math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>

# Forward stepwise selection

- Add the most significant variable at each step
  - Variable with the **smallest p-value**,
  - Variable that provides the highest drop in model error
- The stopping rule is satisfied when all remaining variables to consider have a p-value **larger** than some threshold if added to the model.
  - When we reach this state, forward selection will terminate and return a model that only contains variables with p-values  $<$  threshold.

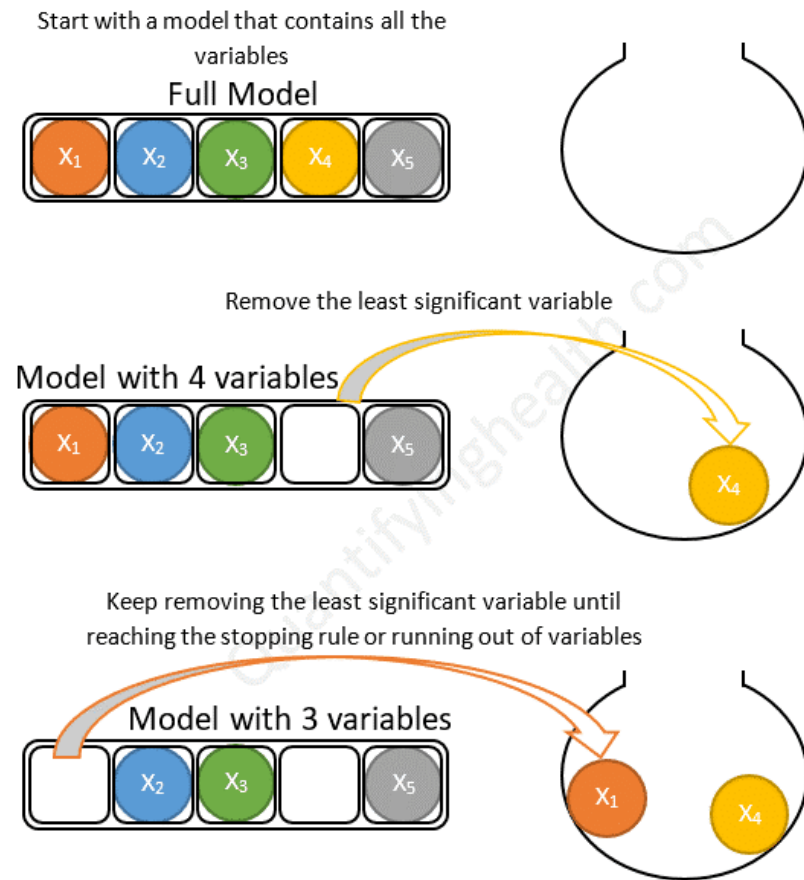
Forward stepwise selection example with 5 variables:



# Backward stepwise elimination

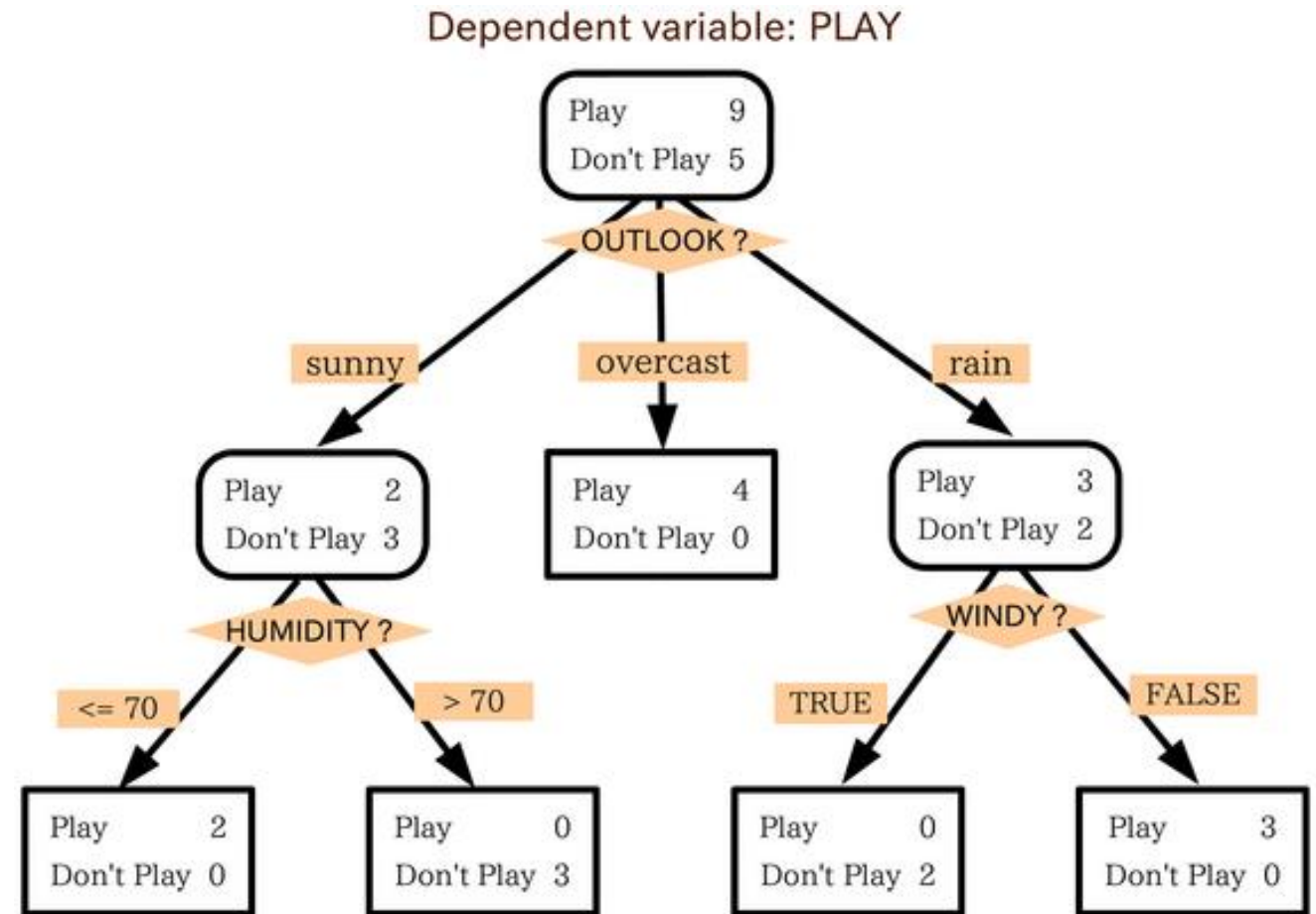
- Remove the least significant variable at each step
  - Variable with the **highest p-value** in the model,
  - Variable when eliminated from the model causes the lowest increase in model error
- The stopping rule is satisfied when all remaining variables in the model have a p-value **smaller** than some pre-specified threshold.
- Possible combination of forward selection and backward elimination
  - selects the best attribute and removes the worst from among the remaining attributes.

Backward stepwise selection example with 5 variables:



# Decision tree induction

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
<b>Rainy</b>	<b>Cool</b>	<b>Normal</b>	<b>True</b>	<b>No</b>
Overcast	Cool	Normal	True	Yes
<b>Sunny</b>	<b>Mild</b>	<b>High</b>	<b>False</b>	<b>No</b>
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
<b>Rainy</b>	<b>Mild</b>	<b>High</b>	<b>True</b>	<b>No</b>



# Feature creation

- Feature creation involves deriving new features from other existing ones.
- Well-conceived new features can sometimes capture the important information in a dataset much more effectively than the original features
- New features can be created by simple mathematical operations such as aggregations to obtain the mean, median, mode, sum, or difference and even product of two values.

# Feature creation

	Candy Variety	Date and Time	Day	Length	Breadth	Price
0	Chocolate Hearts	2020-02-09 14:05:00	Sunday	3.0	2.0	7.5
1	Sour Jelly	2020-10-24 18:00:00	Saturday	3.5	2.0	7.6
2	Candy Canes	2020-12-18 20:13:00	Friday	3.5	2.5	8.0
3	Sour Jelly	2020-10-25 10:00:00	Sunday	3.5	2.0	7.6
4	Fruit Drops	2020-10-18 15:46:00	Sunday	5.0	3.0	9.0

	Candy Variety	Date	Weekend	Price	Size
0	Chocolate Hearts	2020-02-09	1	7.5	6.00
1	Sour Jelly	2020-10-24	1	7.6	7.00
2	Candy Canes	2020-12-18	0	8.0	8.75
3	Sour Jelly	2020-10-25	1	7.6	7.00
4	Fruit Drops	2020-10-18	1	9.0	15.00