SENSOR DATA

DATA

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	5 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	low	safe
poor	3 yrs	high	risky
poor	5 yrs	low	safe
fair	3 yrs	high	safe

x and y values are known

MISSING VALUES

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	?	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	high	safe
poor	?	high	risky
poor	5 yrs	low	safe
fair	?	high	safe

Unknown values

MISSING VALUES IMPACT

Missing values impact both training and prediction

1. Training data: unknown values
2. Prediction: input for prediction has unknown values

MISSING VALUES IMPACT

Training data: "unknown" values

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young		false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true		good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false		No

MISSING VALUES IMPACT

Prediction: input at prediction time with "unknown" values

Age	Has_Job	Own_house	Credit-Rating	Class
young	false		good	?

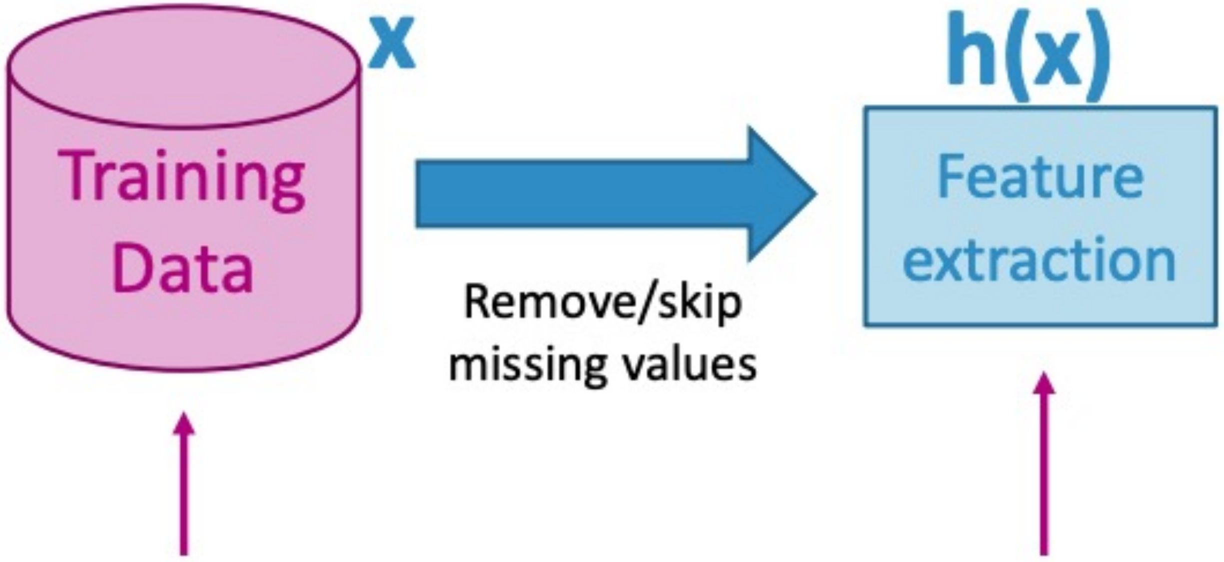
HANDLING MISSING VALUES

Strategy 1: Purification by skipping

PURIFICATION BY SKIPPING / REMOVING

ID	Age	Has_Job	Own_House	Credit_Rating	Class
~~1~~	~~young~~		~~false~~	~~fair~~	~~No~~
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
~~8~~	~~middle~~	~~true~~		~~good~~	~~Yes~~
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
~~15~~	~~old~~	~~false~~	~~false~~		~~No~~

PURIFICATION BY SKIPPING / REMOVING



Original data with missing values

Data without any missing values

THE CHALLENGE WITH SKIPPING / REVOMING

ID	Age	Has_Job	Own_House	Credit_Rating	Class
~~1~~	~~young~~		~~false~~	~~fair~~	~~No~~
2	young	false	false	good	No
~~3~~	~~young~~	~~true~~	~~false~~		~~Yes~~
4	young	true	true	fair	Yes
~~5~~	~~young~~		~~false~~	~~fair~~	~~No~~
6	middle	false	false	fair	No
~~7~~	~~middle~~		~~false~~	~~good~~	~~No~~
~~8~~	~~middle~~			~~good~~	~~Yes~~
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
~~11~~	~~old~~		~~true~~	~~excellent~~	~~Yes~~
12	old	false	true	good	Yes
~~13~~	~~old~~		~~false~~	~~good~~	~~Yes~~
~~14~~	~~old~~		~~false~~	~~excellent~~	~~Yes~~
~~15~~	~~old~~	~~false~~	~~false~~		~~No~~

THE CHALLENGE WITH SKIPPING / REVOMING

ID	Age	Has_Job	Own_House	Credit_Rating	Class
~~1~~	~~young~~		~~false~~	~~fair~~	~~No~~
2	young	false	false	good	No
~~3~~	~~young~~	~~true~~	~~false~~		~~Yes~~
4	young	true	true	fair	Yes
~~5~~	~~young~~		~~false~~	~~fair~~	~~No~~
6	middle	false	false	fair	No
~~7~~	~~middle~~		~~false~~	~~good~~	~~No~~
~~8~~	~~middle~~			~~good~~	~~Yes~~
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
~~11~~	~~old~~		~~true~~	~~excellent~~	~~Yes~~
12	old	false	true	good	Yes
~~13~~	~~old~~		~~false~~	~~good~~	~~Yes~~
~~14~~	~~old~~		~~false~~	~~excellent~~	~~Yes~~
15	old	false	false	fair	No

Warning: more than 50% of the data are removed!

THE CHALLENGE WITH SKIPPING / REMOVING

Idea 2: Skip features with many missing values

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young		false	fair	No
2	young	false	false	good	No
3	young	true	false		Yes
4	young	true	true	fair	Yes
5	young		false	fair	No
6	middle	false	false	fair	No
7	middle		false	good	No
8	middle			good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old		true	excellent	Yes
12	old	false	true	good	Yes
13	old		false	good	Yes
14	old		false	excellent	Yes
15	old	false	false		No

THE CHALLENGE WITH SKIPPING / REMOVING

Strategy 1: Skip data points with a missing value
- make sure only a few points are skipped

Strategy 2: Skip features with many missing values
- make sure only a few features are skipped

SKIPPING / REMOVING MISSING VALUES: PROS AND CONS

Pros:

- Easy to understand and implement
- Applied to all machine learning model

Cons:

- Removing data points and features may take off some important information
- Unclear when it's better to remove data points or features
- Doesn't help if data is missing at prediction part

HANDLING MISSING VALUES

Strategy 2: Purification by imputing

MAIN DRAWBACK OF SKIPPING METHOD

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young		false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true		good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false		No

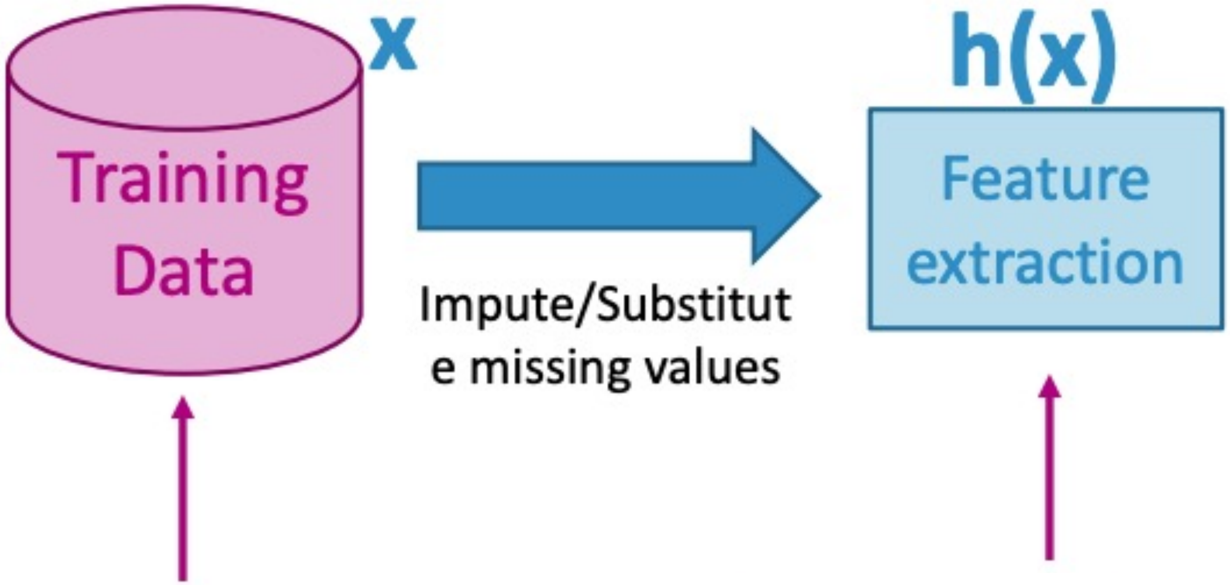
Data is precious.
Do not throw it away.

CAN WE KEEP ALL THE DATA?

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young		false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true		good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false		No

Use other data point in the column to “guess” the “missing part”.

IDEA: PURIFICATION BY IMPUTING



Original data with missing values

Same number of data points

IDEA: PURIFICATION BY IMPUTING

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Fill in each missing value with a calculated guess

EXAMPLE: REPLACE WITH THE MOST COMMON VALUE

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true		good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false		No

Fill in each missing value with a calculated guess

COMMON (SIMPLE) RULES FOR IMPUTING

Impute each feature with missing values:

1. Categorical features: Most popular value of non-missing
2. Numerical features: Average or median value of non-missing

MISSING VALUE IMPUTATION: PROS AND CONS

Pros

- Easy to understand and implement
- works for all machine learning models
  (logistic regression, decision trees, …)
- works for missing values in the prediction part
  use the same imputation rules

Cons

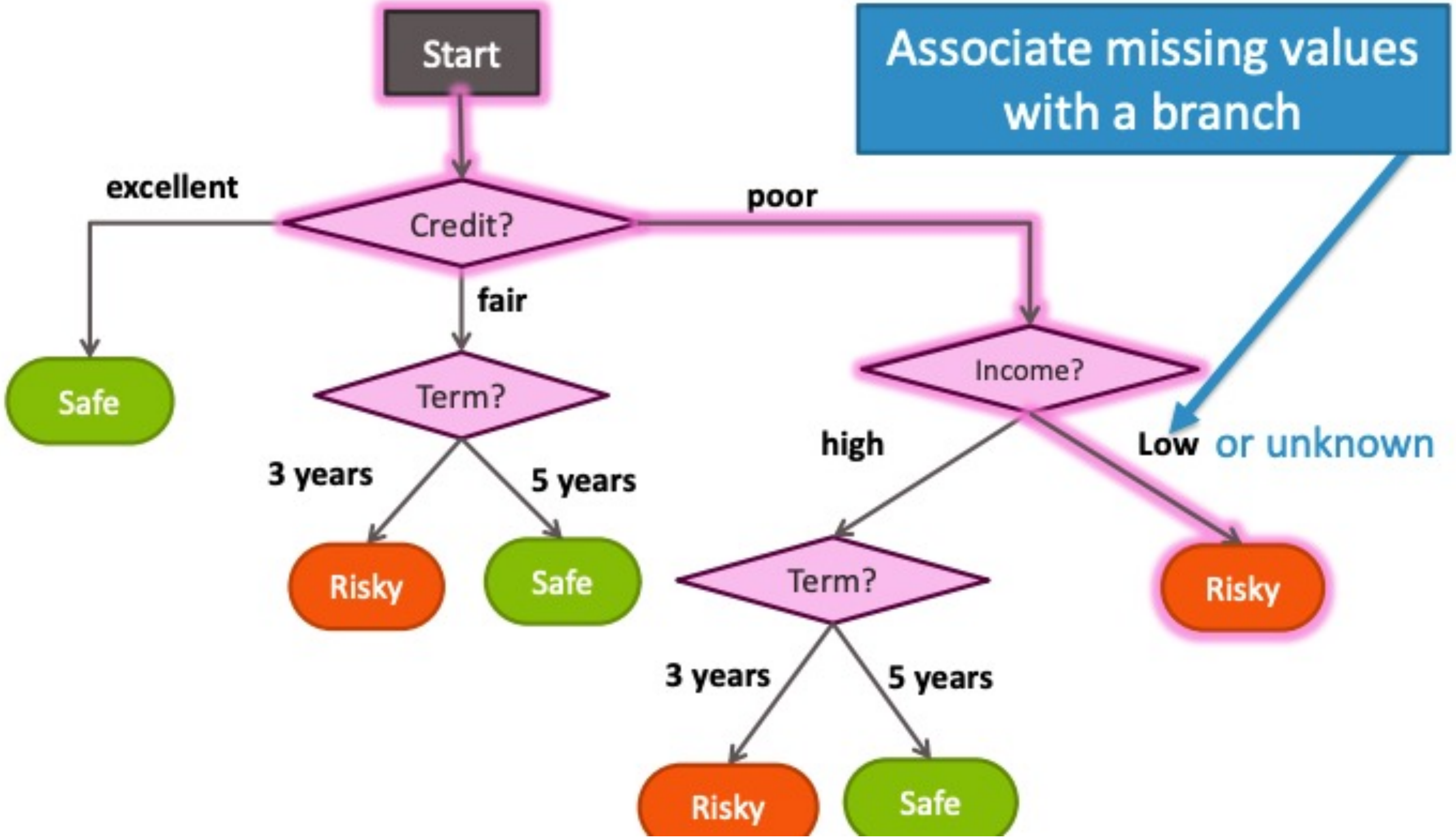
- May have systematic errors

Example: a feature is missing in the entire dataset in one place but is not missing in another dataset.

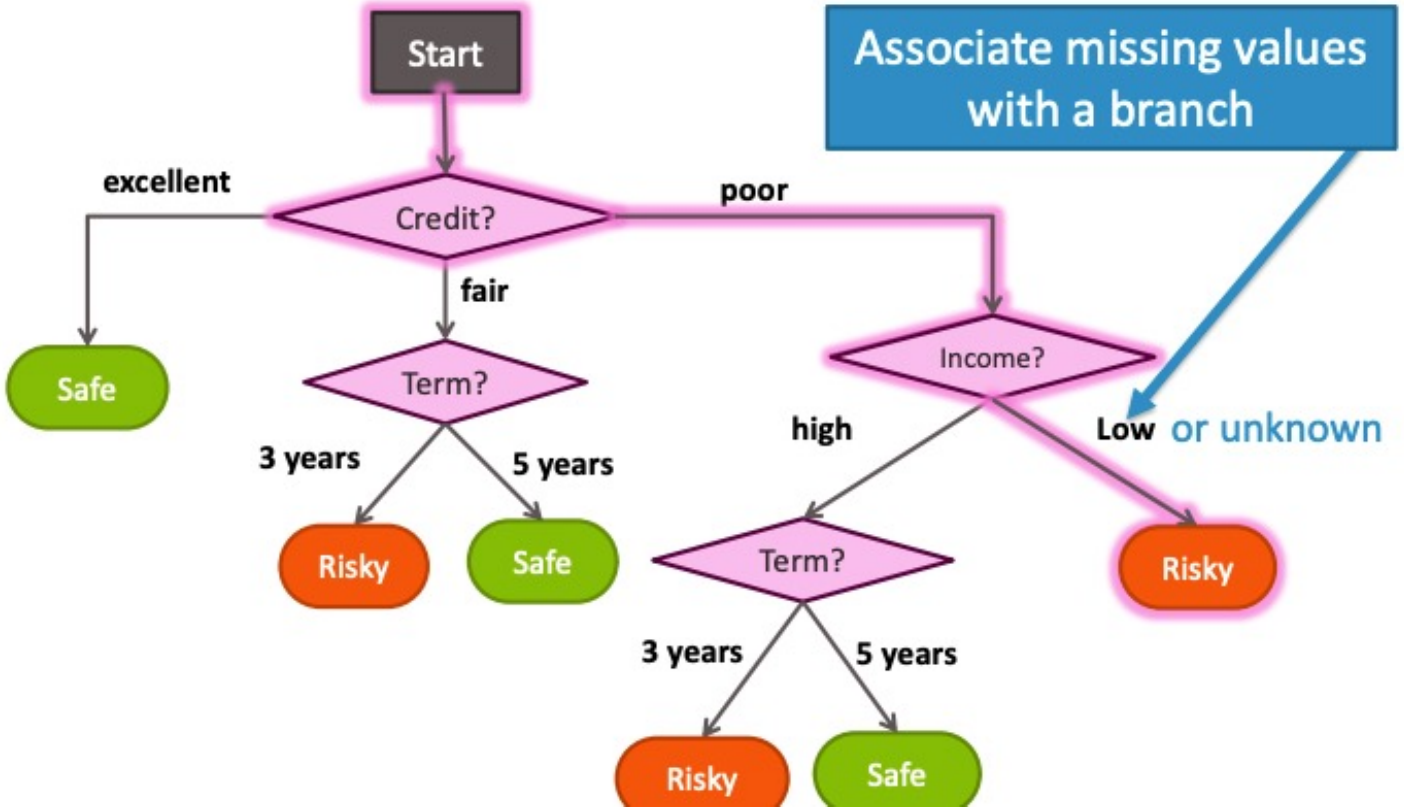
HANDLING MISSING VALUES

Strategy 3: Adapt learning algorithm to be robust to missing values

HANDLING MISSING DATA



HANDLING MISSING DATA



Every decision node includes choice of response to missing values

FEATURE SPLIT SELECTION WITH MISSING DATA

Pros

- works in both training and prediction parts
- More accurate predictions

Cons

- modify learning algorithms
  (simple for decision trees)

SUMMARY OF HANDLING MISSING VALUES

WHAT YOU CAN DO NOW…

Describe common ways to handling missing data:

1. Skip all data points (rows) with any missing values
2. Skip features (columns) with many missing values
3. Impute missing values
4. Modify learning algorithm (decision trees)

DATA PREPROCESSING

- Data Cleaning (missing values)
- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Integration
- Data Reduction (PCA)

DATA QUALITY: WHY PREPROCESS THE DATA?

- Measures for data quality: A multidimensional view
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability

MAJOR TASKS IN DATA PREPROCESSING

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration (data engineer)**
  - Integration of multiple databases (sql), data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression

NOISY DATA

- Noise: random error or variance
- Incorrect attribute values may be due to
  - data collection instrument failures
  - data transmission problems
  - technology limitations
- Other data problems which require data cleaning
  - Duplicate, incomplete, inconsistent

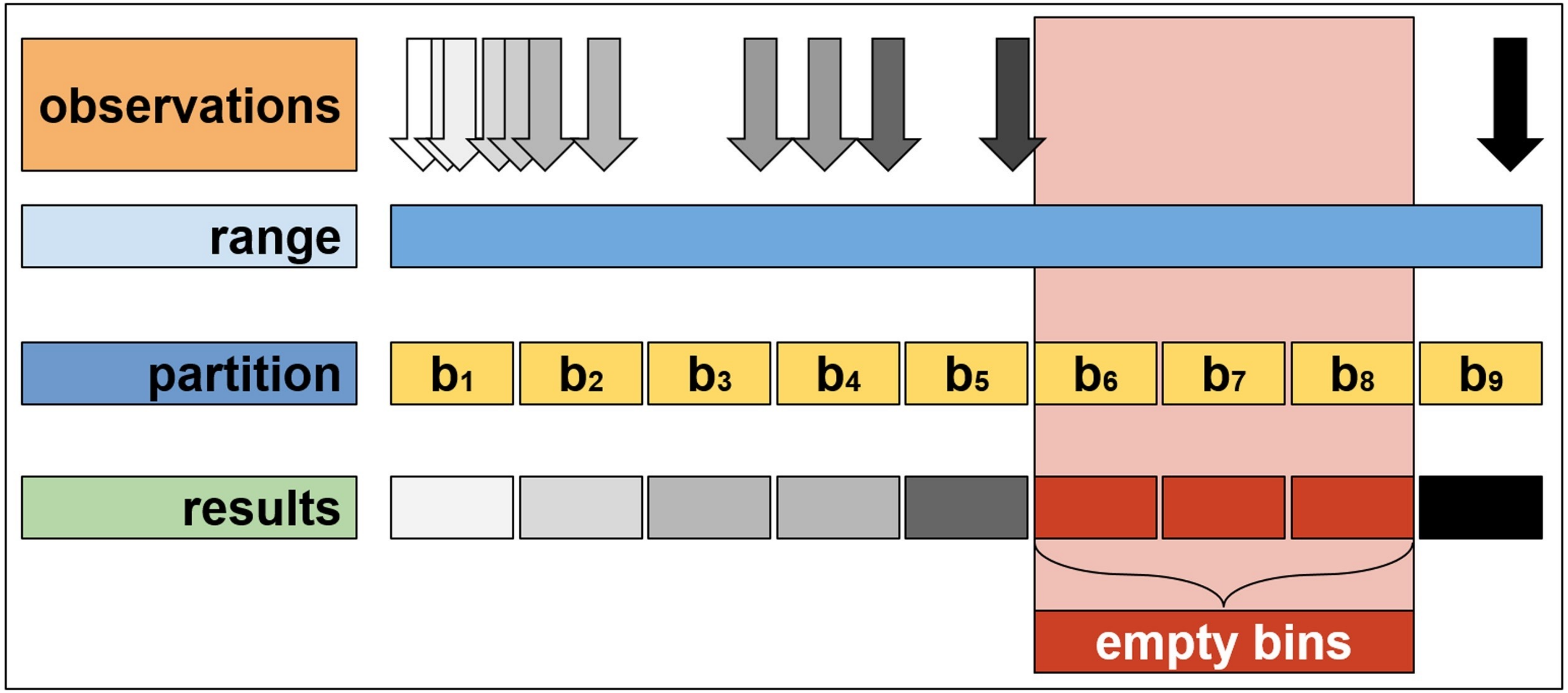
HOW TO HANDLE NOISY DATA?

- Binning
- Regression – supervised learning
- Clustering (unsupervised learning)
- Combined computer and human inspection

HOW TO HANDLE NOISY DATA?

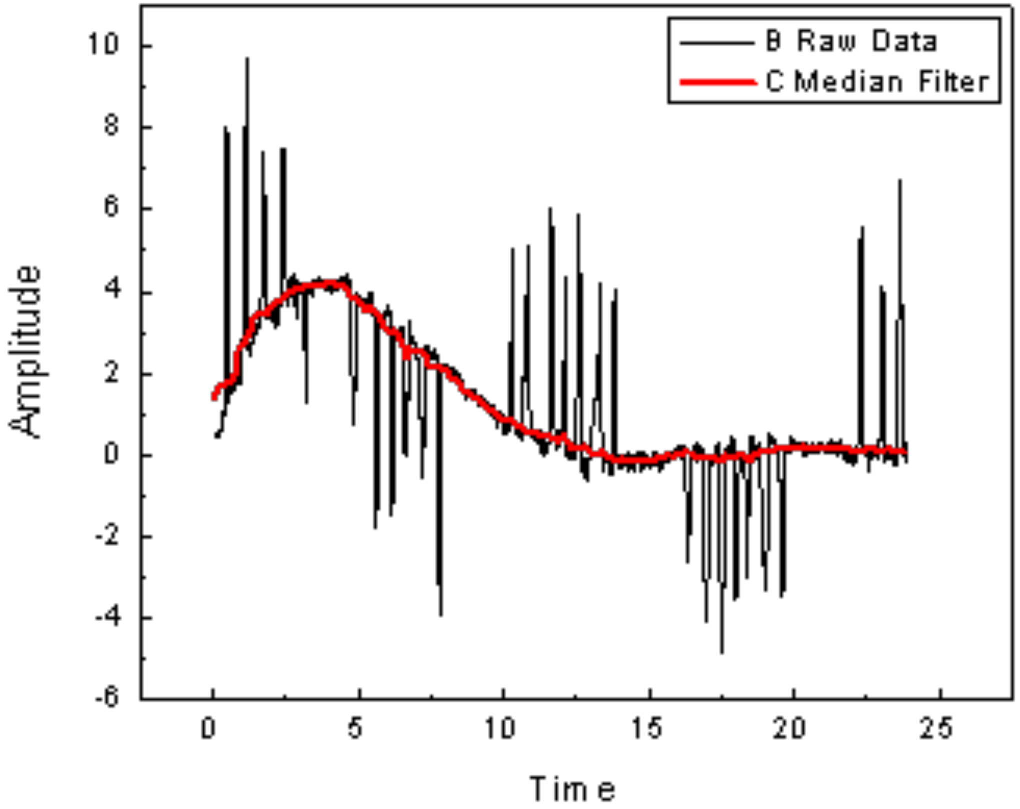
- Binning (numerical – data engineer)
  - first sort data
  - partition sorted data into (equal-frequency) bins
  - smooth by bin *means*, *median*, or *boundaries* (e.g, clean jitters)

HOW TO HANDLE NOISY DATA?



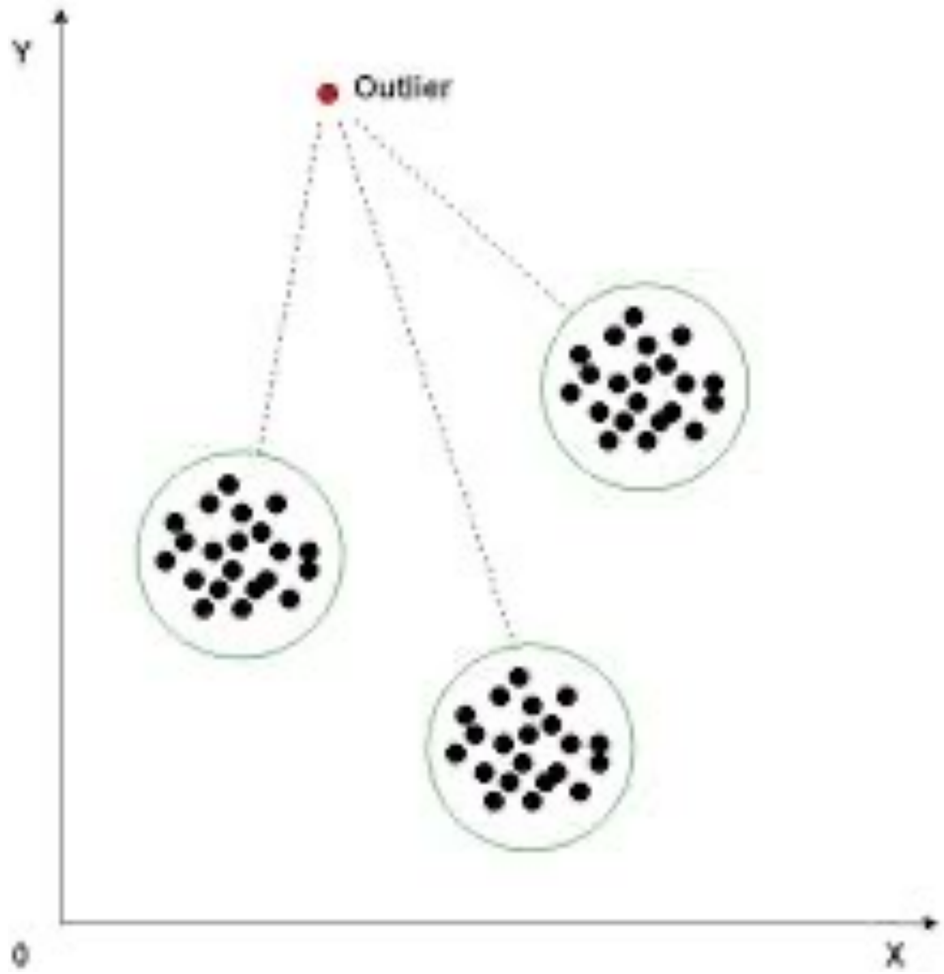
HOW TO HANDLE NOISY DATA?

- Regression
- smooth by fitting the data into regression functions



HOW TO HANDLE NOISY DATA?

- Clustering (unsupervised learning)
  - detect and remove outliers



HOW TO HANDLE NOISY DATA?

- Combined computer and human inspection
- (human in the loop ⇔ combine domain experts' perspectives)
  - detect suspicious values and check by human (e.g., deal with possible outliers)