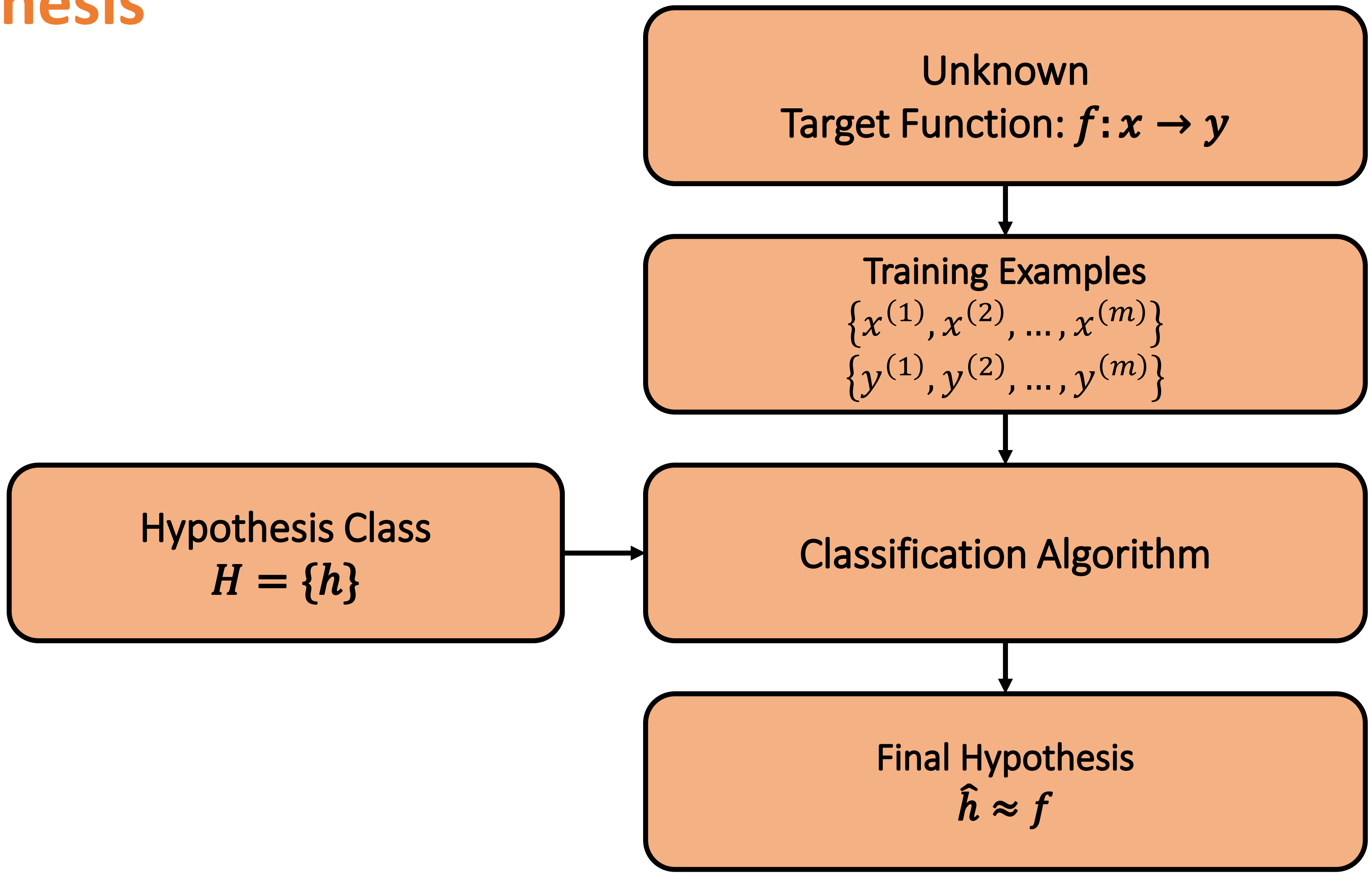
Machine Learning in Healthcare

#L02-Data exploration and preprocessing

Technion-IIT, Haifa, Israel

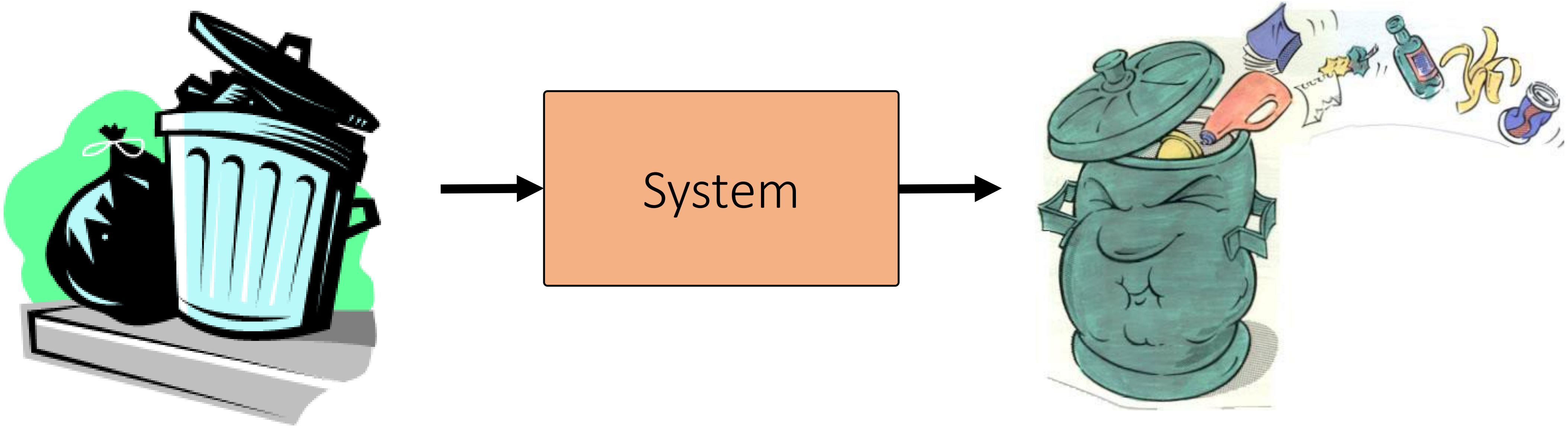
Asst. Prof. Joachim Behar
Biomedical Engineering Faculty, Technion-IIT
Artificial intelligence in medicine laboratory (AIMLab.)
https://aim-lab.github.io/
Twitter: @lab_aim

Hypothesis


Unknown
Target Function: $f: x \rightarrow y$
Training Examples
$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
$\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$
Hypothesis Class
$H = \{h\}$
Classification Algorithm
Final Hypothesis
$\hat{h} \approx f$

The importance of preprocessing

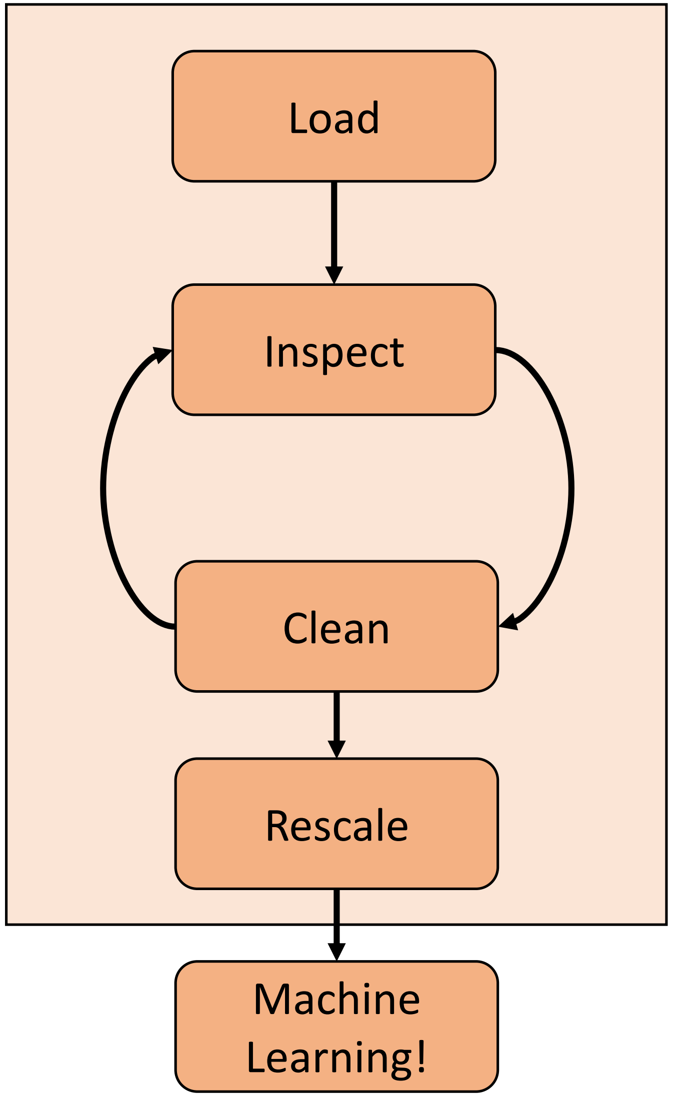
- Garbage in ➔ Garbage out.
- Poor data will lead to bad predictions.



Topics covered

- Load: data types,
- Inspect: exploratory data analysis,
- Clean: handle abnormalities,
- Rescale: rescale the data.

Data preprocessing

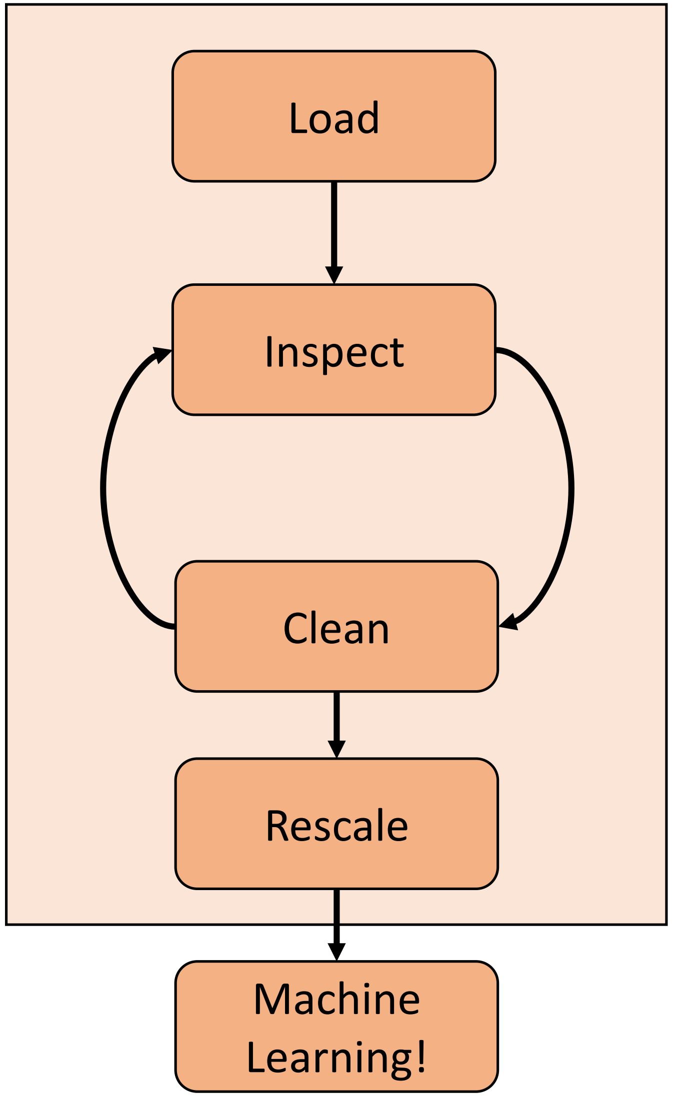


Data types

Topics covered

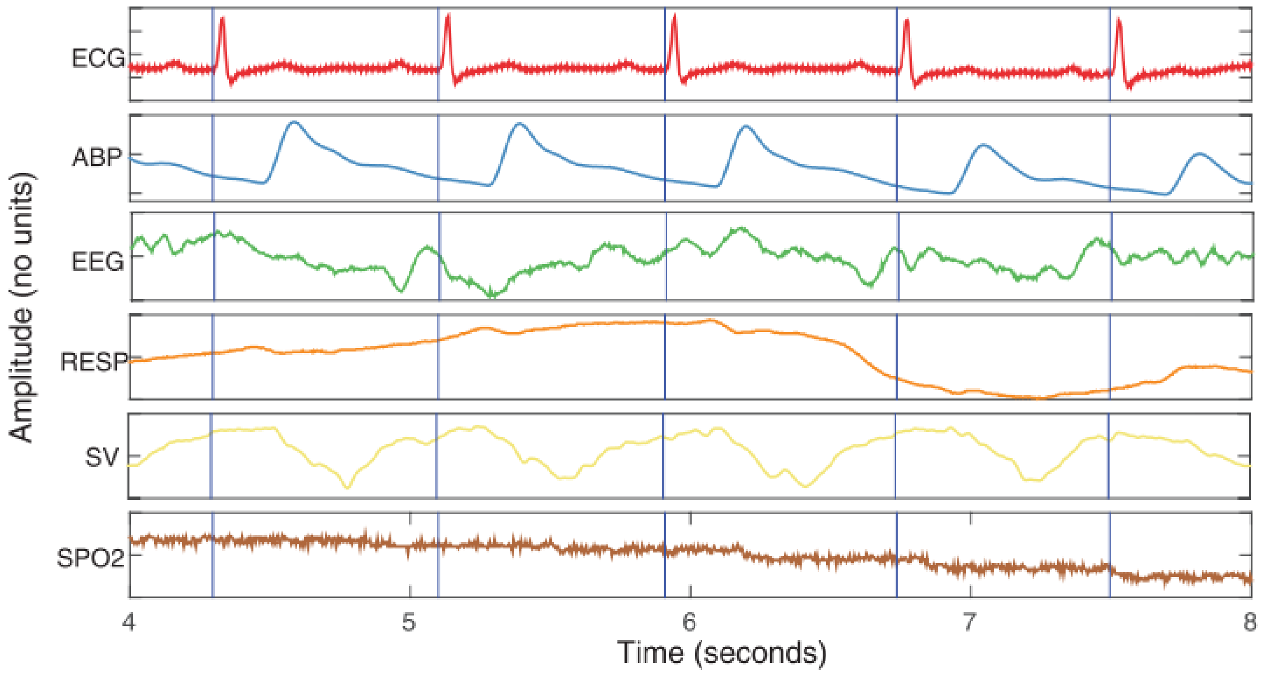
- Load: data types,
  - Sources of data,
  - Types of data.
- Inspect: exploratory data analysis.
- Clean: handle abnormalities.
- Rescale: rescale the data.

Data preprocessing



AIMLab.

Sources of data



Johnson AW, Behar J. et al. Phys. Meas. 2014

header

header <1x226 cell>

	1	2	3	4	5	6	7	8	9
1	ALPFirst	ALPHighest	ALPLast	ALPLowest	ALPMedian	ALPNumRe...	ALTFirst	ALTHighest	ALTLas
2									

X

X <4000x226 double>

	1	2	3	4	5	6	7	8
1	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
3	127	127	105	105	127	2	91	91
4	105	105	105	105	105	1	12	12
5	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
6	101	101	101	101	101	1	45	60
7	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
8	47	47	47	47	47	1	46	46
9	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
10	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
11	402	402	402	402	402	1	36	36
12	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
13	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
14	19	19	19	19	19	1	15	15
15	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
16	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
17	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN

PhysioNet ICU dataset

Types of data

Type

I. Numerical (double)

II. Numerical (int)

III. Boolean — True/false.

IV. Categorical — Can take a limited and fixed number of possible values.

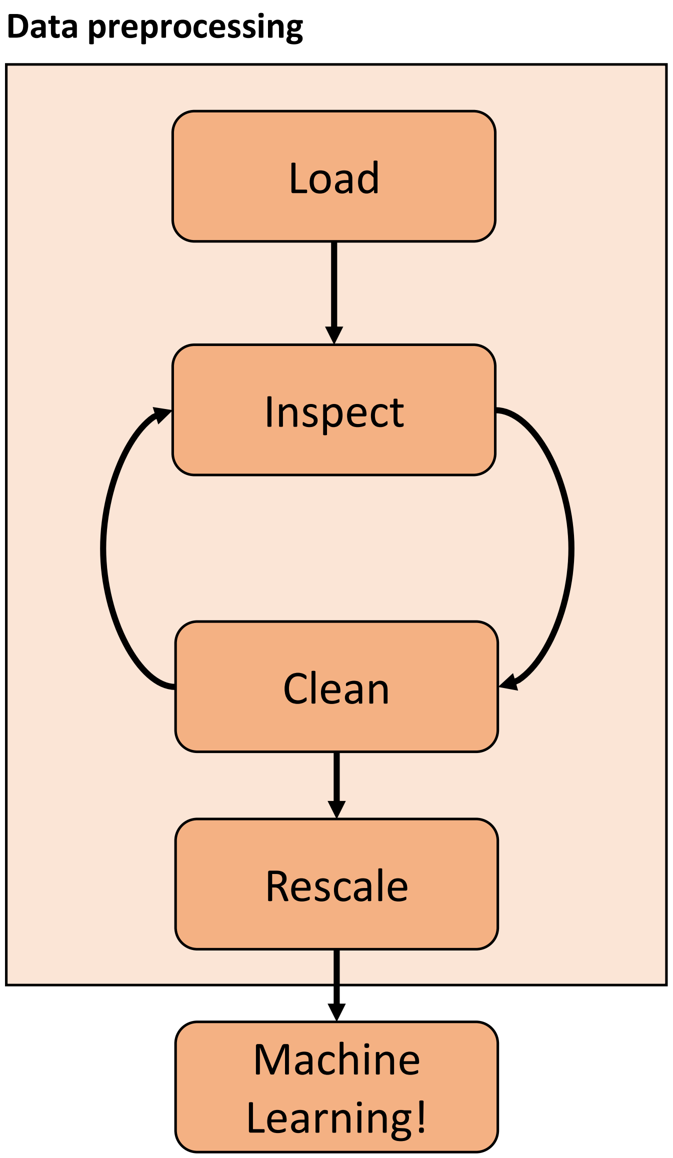
V. Ordinal — Categorical but the variables have natural ordered categories and the distance between the categories is not known.

VI. Others

Exploratory data analysis

Topics covered

- Load: data types,
- Inspect: exploratory data analysis,
  - Summary statistics,
  - Data visualization.
- Clean: handle abnormalities,
- Rescale: rescale the data.



Exploratory data analysis

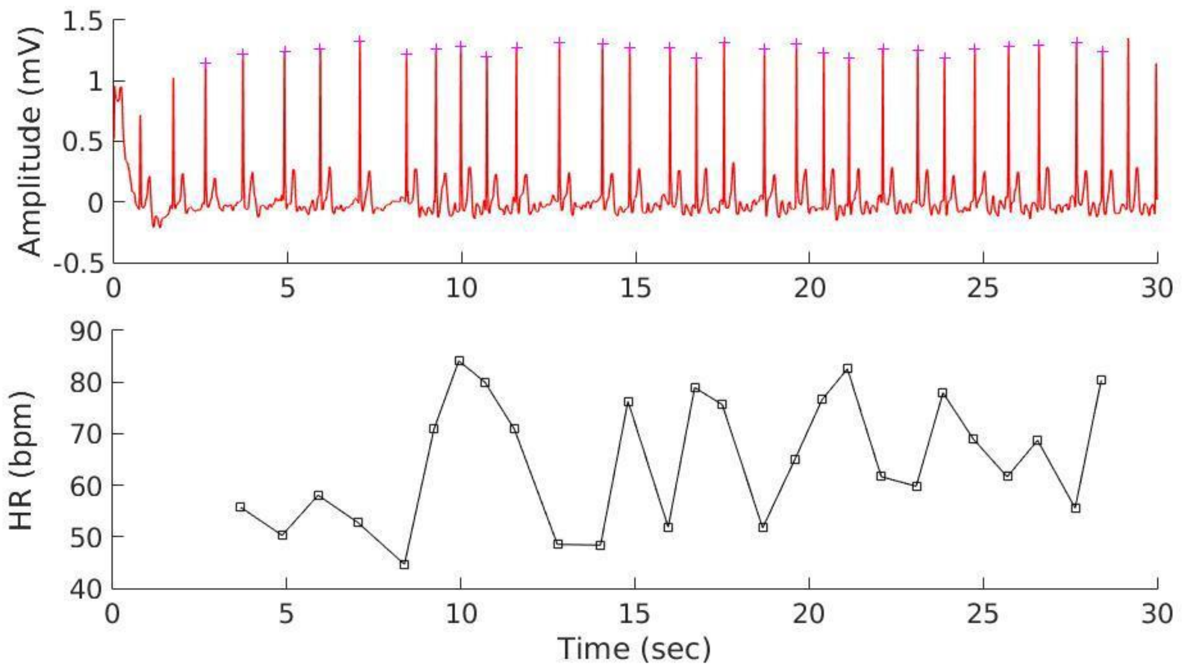
- Descriptive statistics and graphical representation of the data for the purpose of exploring the data.
  - **Data visualization**: summarize the distribution and relationships between variables using visualizations such as charts, plots and graphs.
  - **Summary statistics:** summarize the distribution and relationships between variables using statistical quantities.

Data visualization

- Why do we need visualization?
  - To gain a qualitative understanding of the dataset.
  - It can help identify patterns, corrupt data, outliers etc.
  - Data visualization and exploratory data analysis are fields of research by themselves.
- We will look at popular visualization tools:
  - Line plot,
  - Bar chart,
  - Histogram plot,
  - Boxplot,
  - Scatter plot.

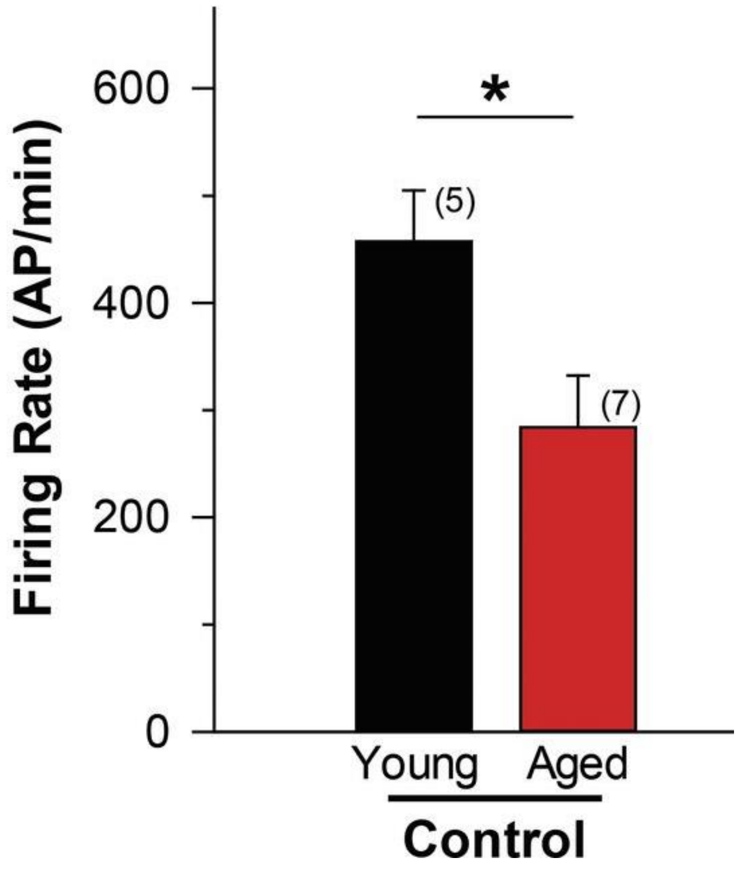
Line plots

- Present observation collected at regular intervals.
- The x-axis represents the regular interval such as time and the y-axis the observations, ordered by the x-axis and connected by a line.



Bar chart

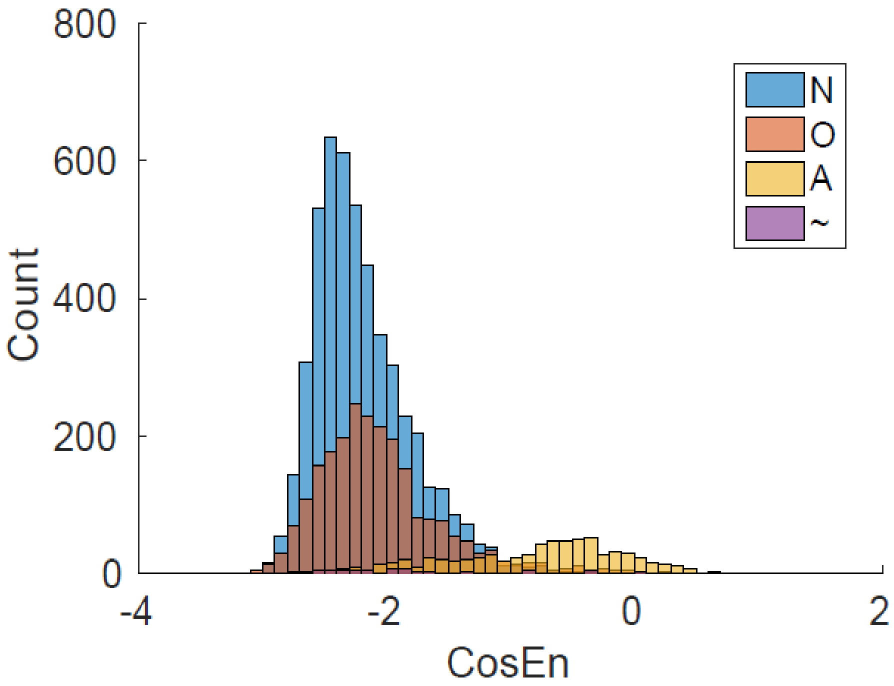
- Present relative quantities for multiple categories.
- The x-axis represents the categories and the y-axis represents the quantity for each category.
- Often with confidence intervals i.e. error bar.



Sharpe et al. *JGP* 2017

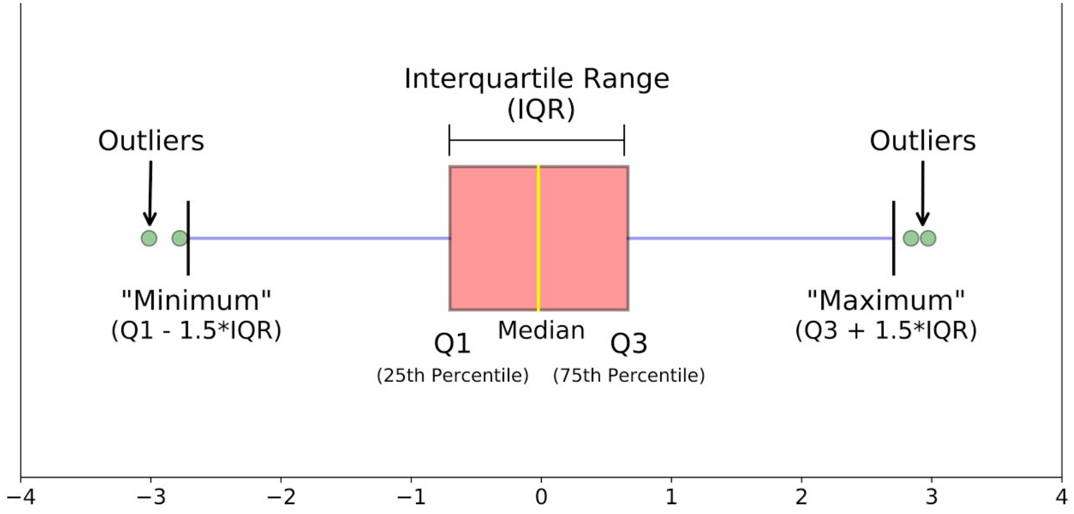
Histogram plots

- Summarize the distribution of a data sample.
- The x-axis represents discrete bins or intervals for the observations and the y-axis represents the frequency or count of the number of observations in the dataset that belong to each bin.



Behar et al. *CinC* 2017

Boxplots

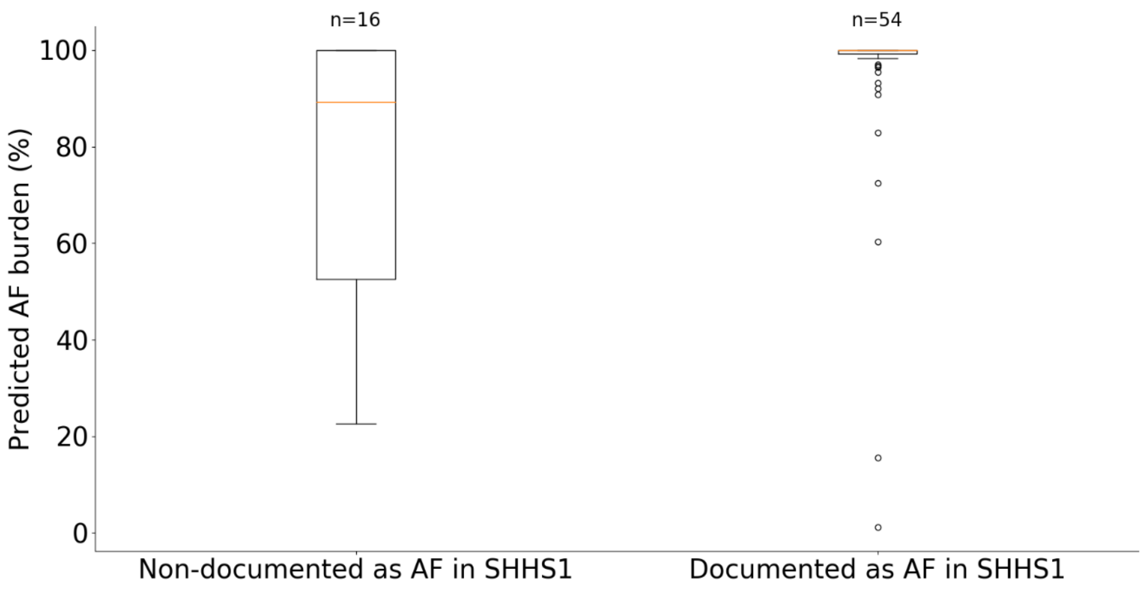


- Summarize the distribution of a data sample.
- The central box of the boxplot summarizes the middle 50% of the dataset. The box starts at the 25th percentile and ends at the 75th percentile.
- The interquartile range (IQR) is computed by the different between the 75th and 25th percentiles.
- Lines called whiskers are drawn extending from both ends of the box with the length of 1.5 x IQR to demonstrate the expected range of sensible values in the distribution.
- Observations outside the whiskers might be outliers and are drawn with small circles.

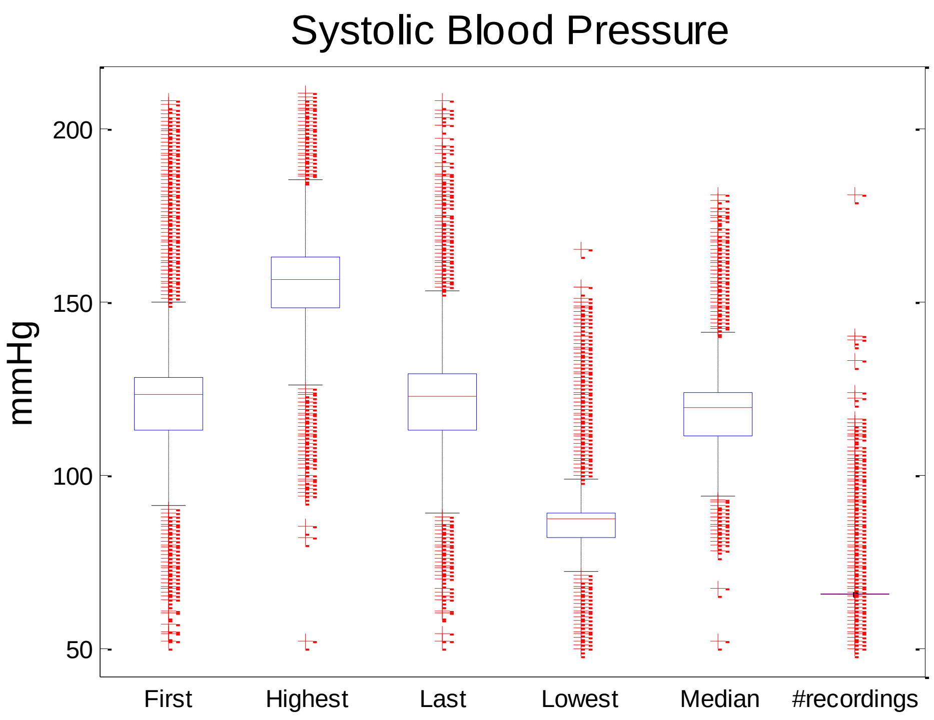
Image: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Boxplots

- Boxplot is a graphical representation of the five number summary statistics.

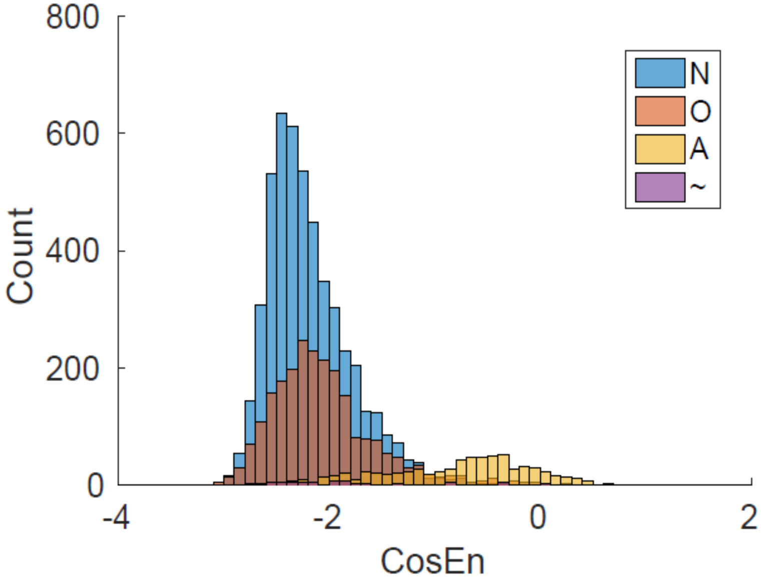


Chocron et al. *Phys. Meas.* 2020



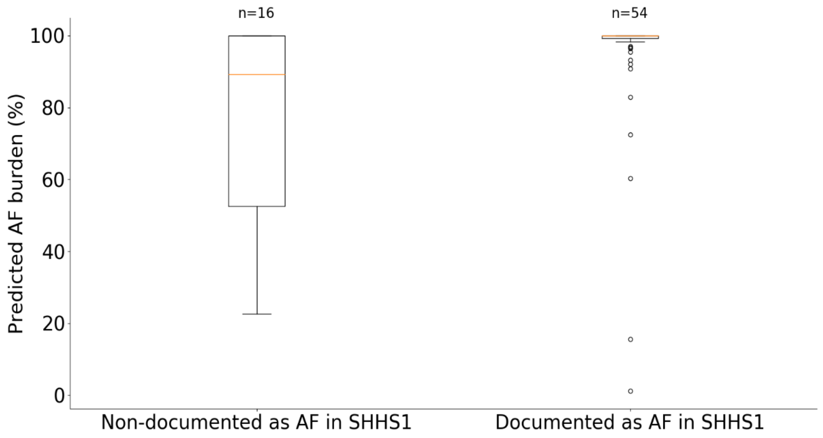
Histogram, boxplot or barplot for my paper?

Histogram

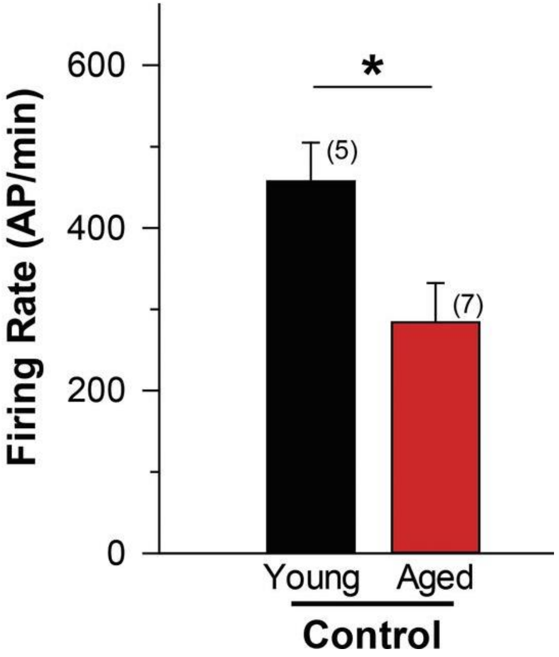


Behar et al. *CinC* 2017

Boxplot



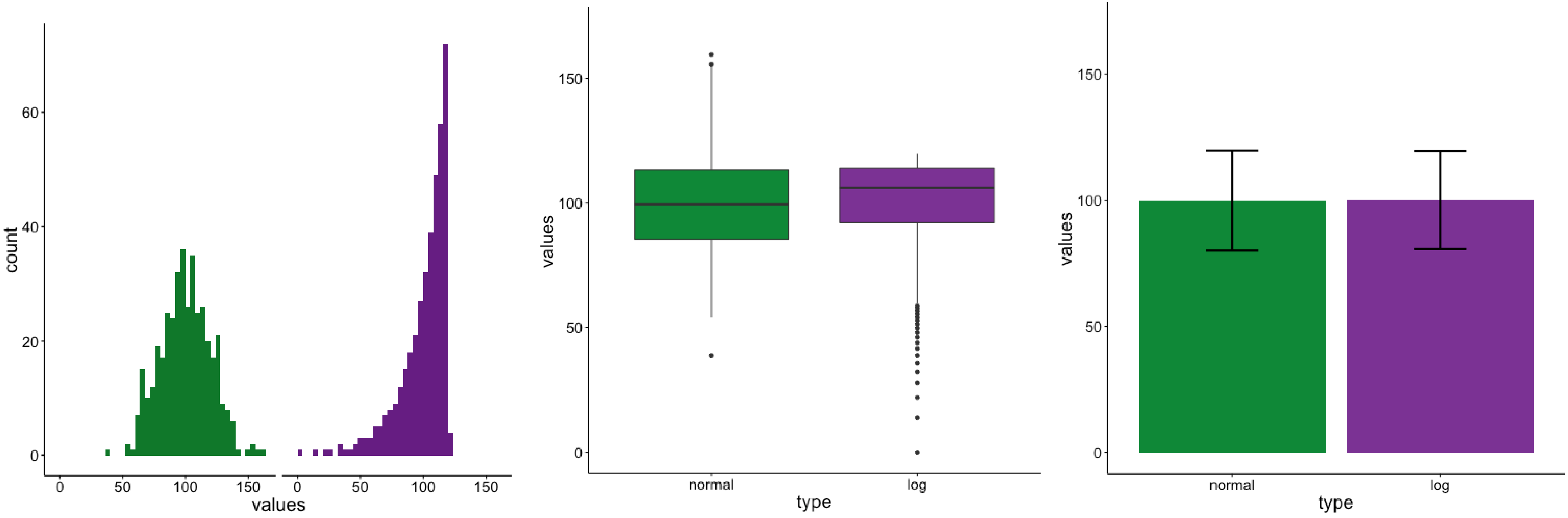
Barplots



Sharpe et al. *JGP* 2017

Note: Histogram, boxplot or barplot for my paper?

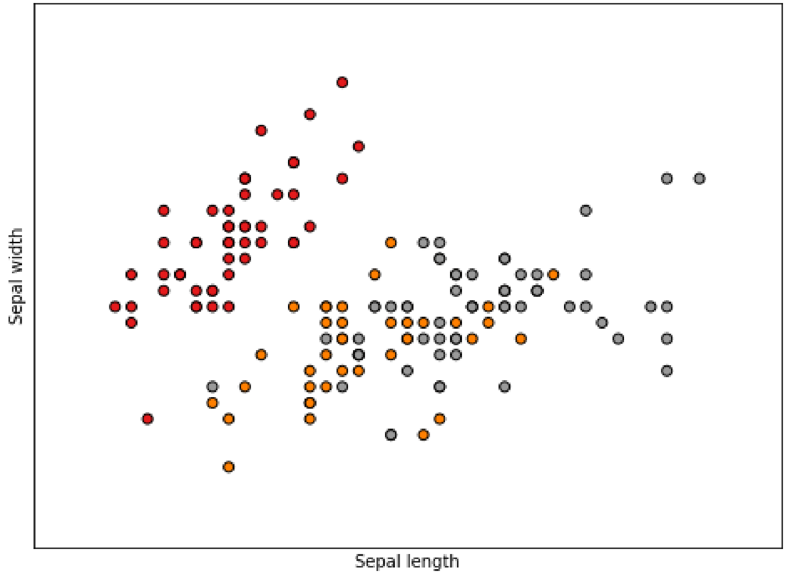
- Figures below: same data, three representations
- Bottom line, prefer histogram and boxplot over bar charts.



https://pagepiccinini.com//stolprab-sv-stolpxob/2016/02/23

Scatter Plot

- Used to summarize the relationship between two paired data samples, e.g. two features of the examples set.
- The x-axis represents observation values for the first sample and the y-axis represents the observation values for the second sample.
- Each point on the scatter plot thus represent a single example.



Violin plots

- Show the probability density of the data at different values, usually smoothed by the kernel density estimator
- Thus they are more informative than the boxplots in that sense.
- It is a useful tool for data representation.

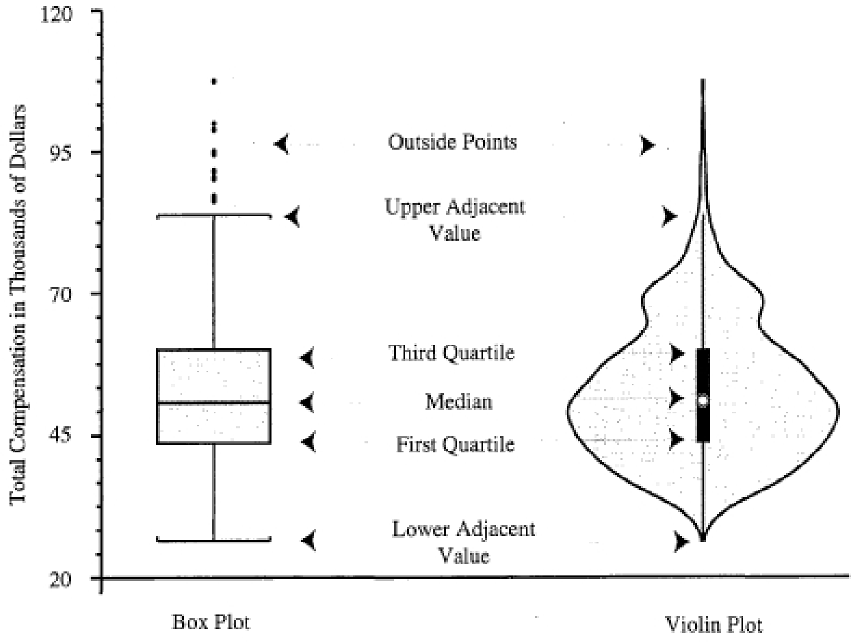


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Image: Hintze, Jerry L., and Ray D. Nelson. "Violin plots: a box plot-density trace synergism." *The American Statistician* 52.2 (1998): 181-184.

AIMLab.

Violin plots

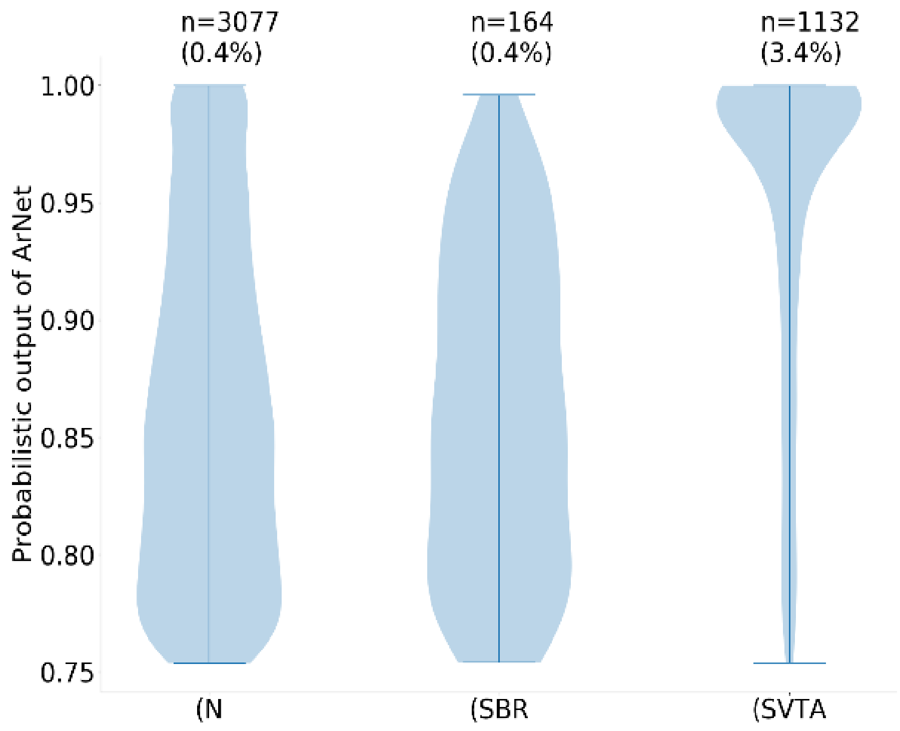


Image: Chocron et al. *Under review IEEE TBME.* 2020

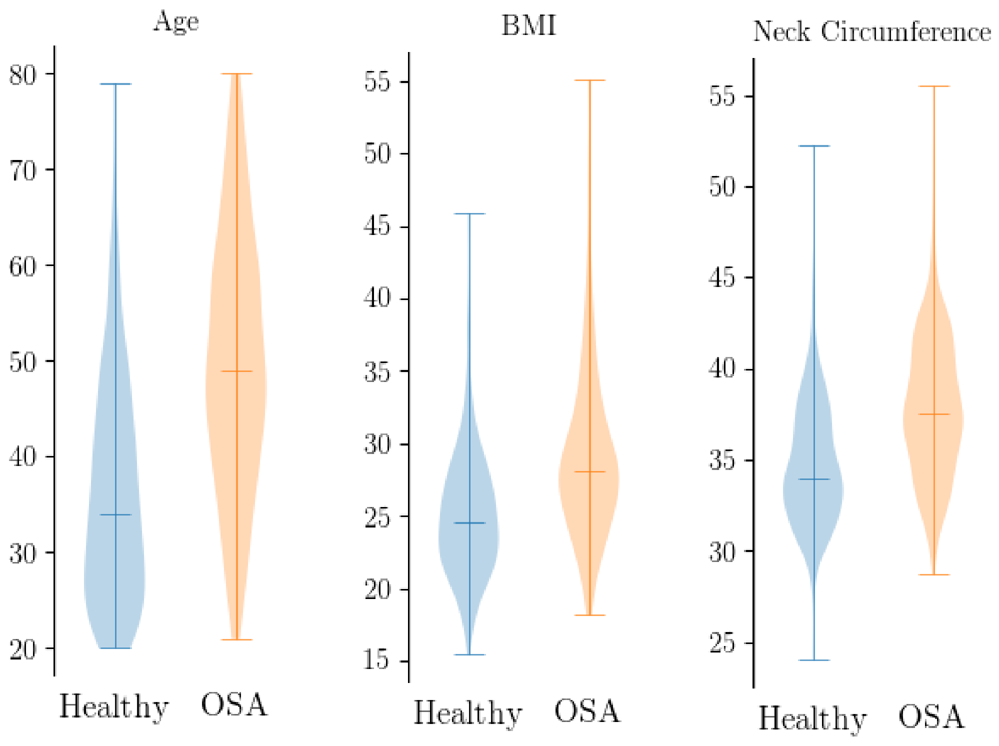
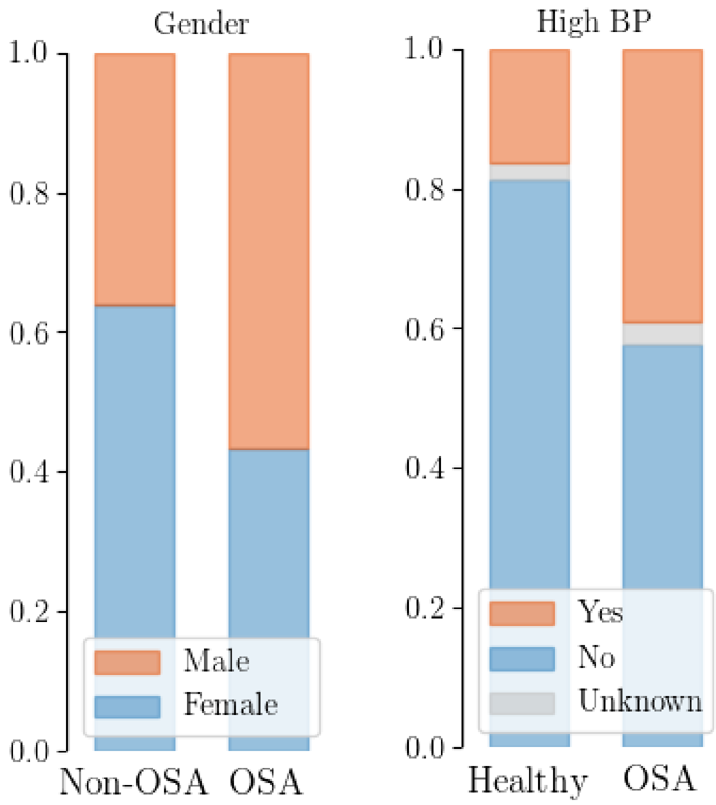
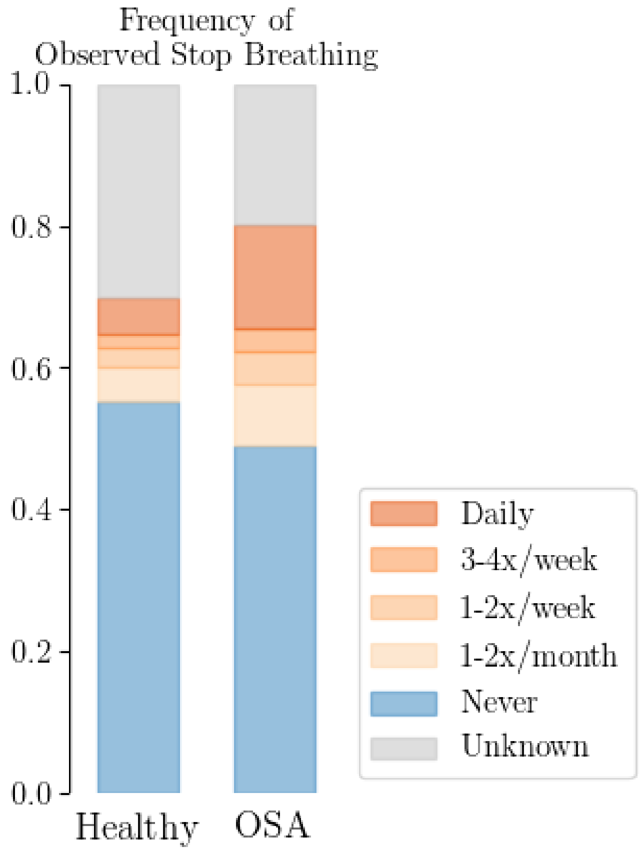


Image: Behar, Palmius et al. *Phys. Meas.* 2020

Binary and ordinal features



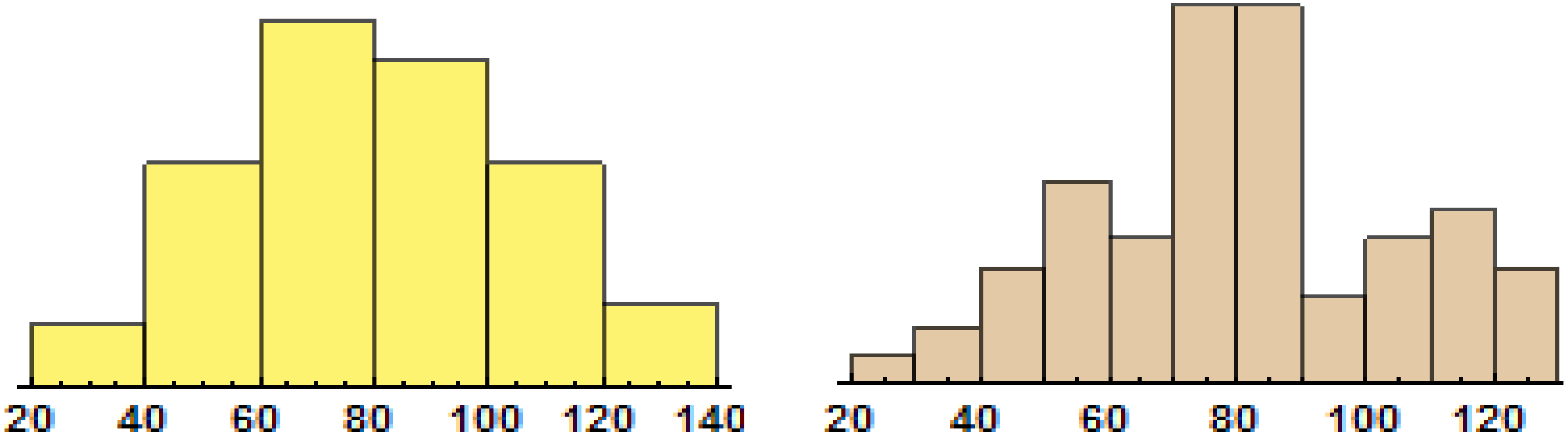


Behar, Palmius et al. *Phys. Meas. 2020*

Note: selecting the number of bins in histograms

- Discretization error.
- Can make data look differently when sample sizes are small.

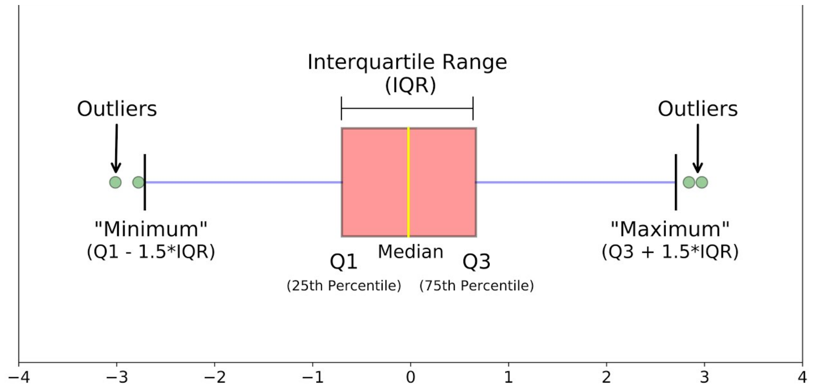
Same data, different number of bins

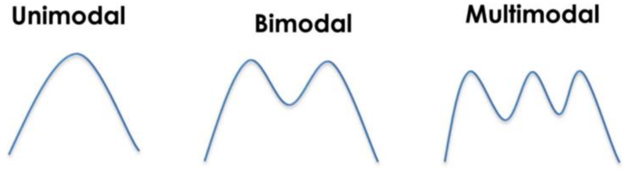


https://pagepiccinini.com//stolprab-sv-stolpxob/2016/02/23

Descriptive statistics

- **Central tendency**: mean, median and mode.
- Measure of **variability/spread**:
  - Range: min, max.
  - Variance and standard deviation.
  - Interquartile range (IQR).
- **Modality:** number of peaks it contains.

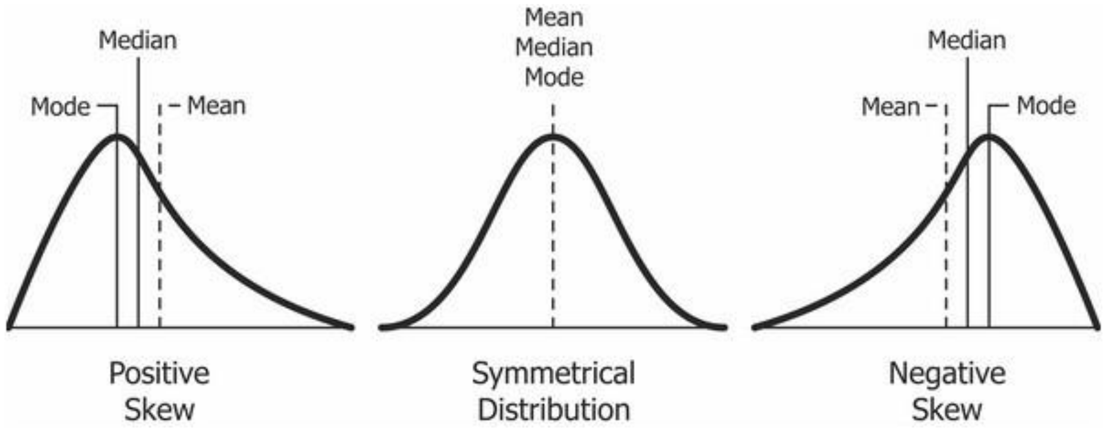




Pictures: https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9

Descriptive statistics

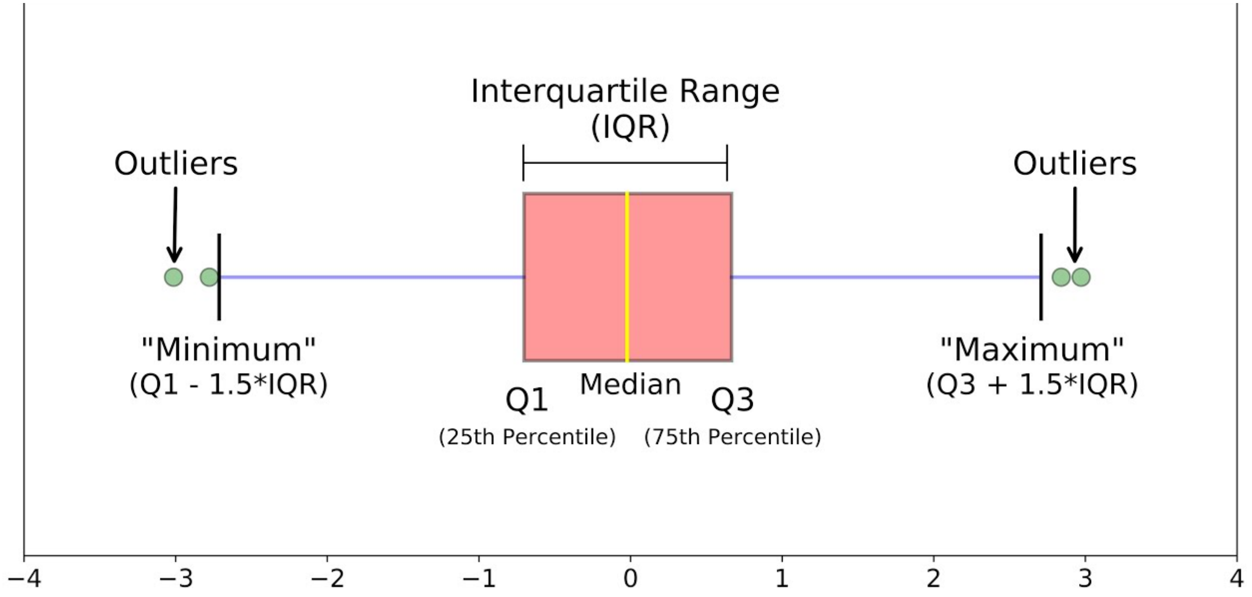
- **Skewness:** measure of the symmetry of a distribution.
- **Kurtosis:** measures whether your dataset is heavy-tailed or light-tailed compared to a normal distribution.



Pictures: https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9

Summary statistics

- Typically you will report the "Five Number Summary statistics": **min, Q1, median, Q3, max**.
- This is complementary to the **boxplot** visualization figure:



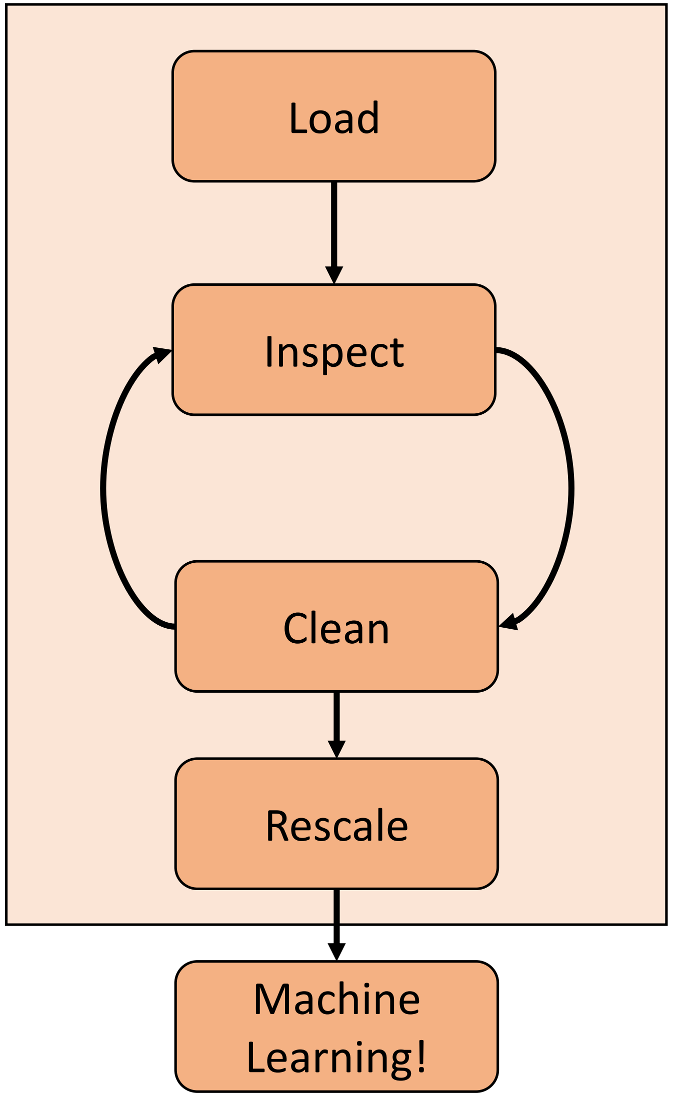
Pictures: https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9

Handling abnormalities

Topics covered

- Load: data types.
- Inspect: exploratory data analysis.
- Clean: handle abnormalities,
  - Missing value,
  - Outliers,
  - Incorrect entries.
- Rescale: rescale the data.

Data preprocessing



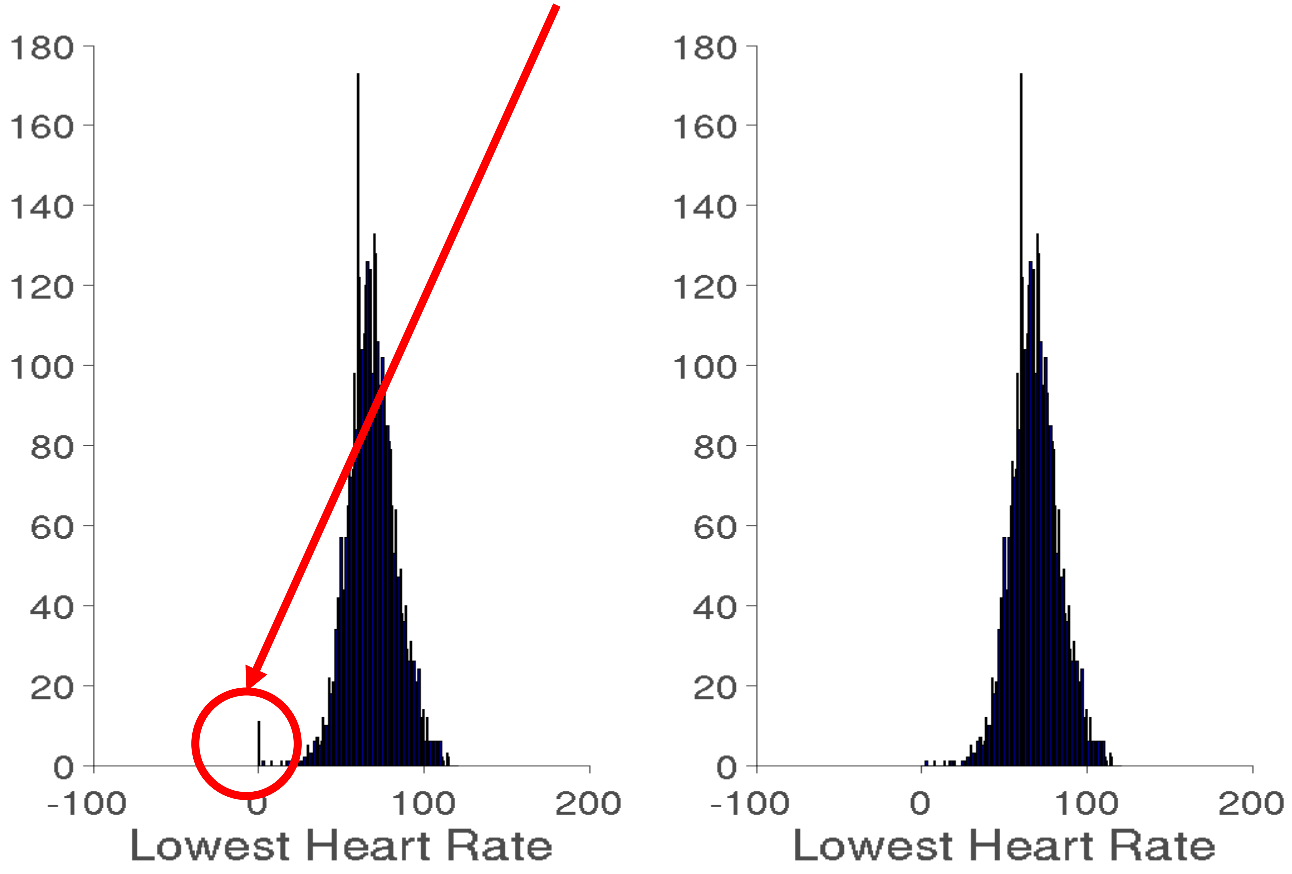
Clean

- Look for the dust!
- Different types of abnormalities:
  - Missing values,
    - Including values which represent missing values,
  - Outliers.
  - Incorrect entries.

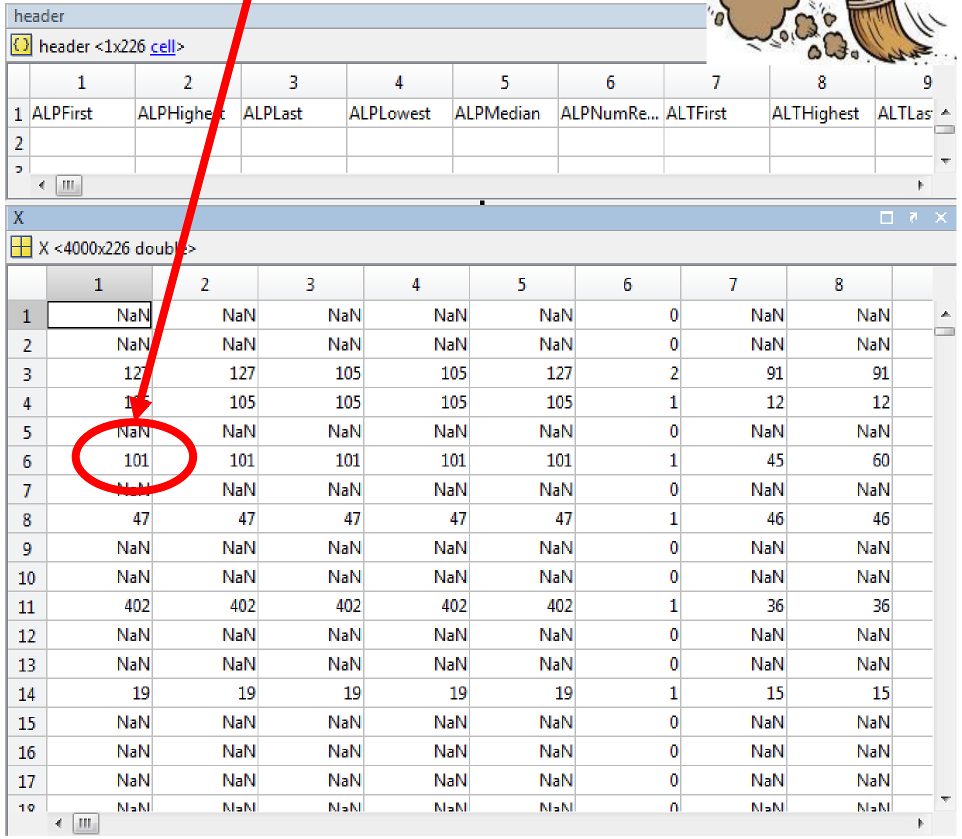
AIMLab.

Look for the dust!

Incorrect entry or missing value



Missing values



header

header <1x226 cell>

	1	2	3	4	5	6	7	8	9
1	ALPFirst	ALPHighe	ALPLast	ALPLowest	ALPMedian	ALPNumRe...	ALTFirst	ALTHighest	ALTLas
2									

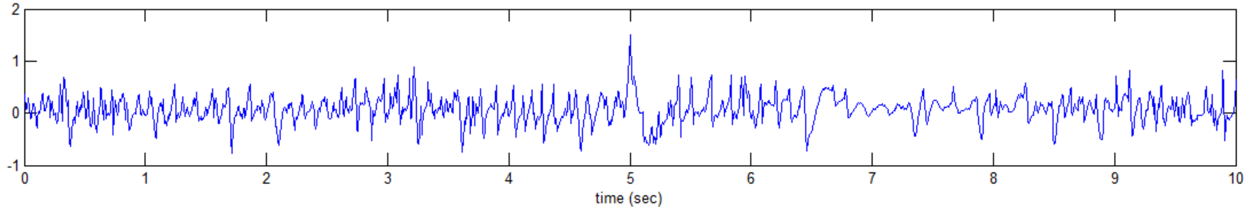
X

X <4000x226 double>

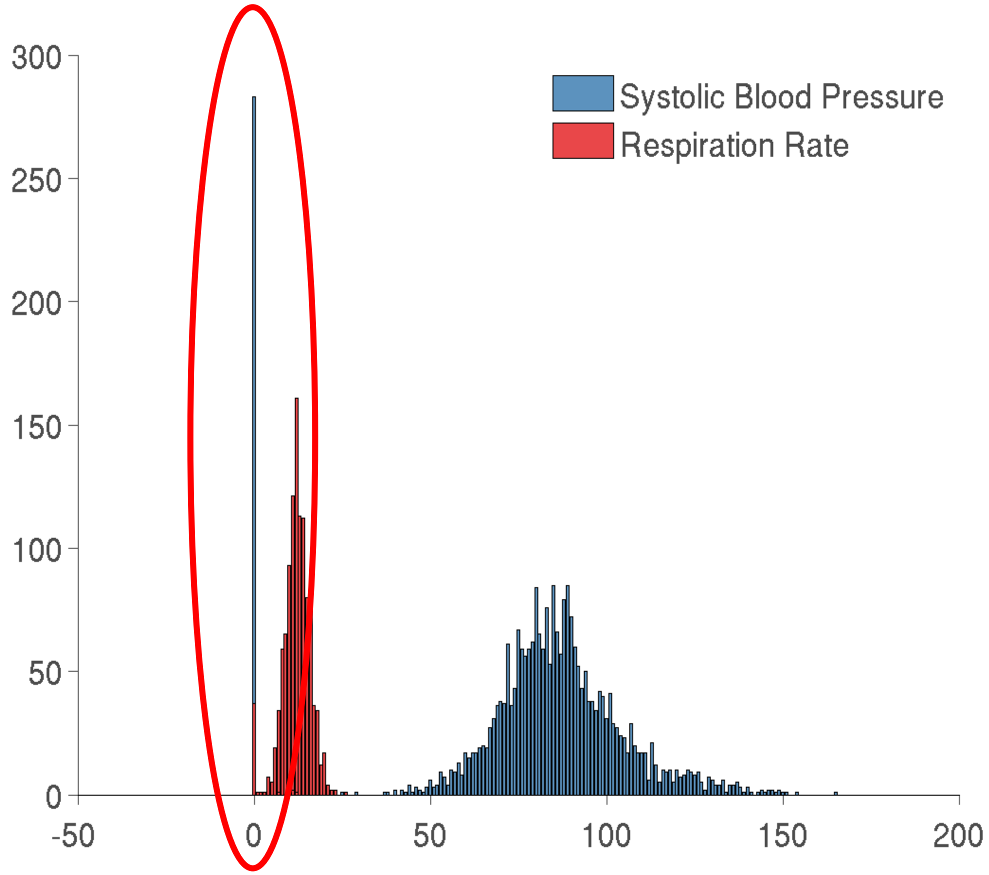
	1	2	3	4	5	6	7	8
1	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
3	[illegible]	127	105	105	127	2	91	91
4	[illegible]	105	105	105	105	1	12	12
5	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
6	101	101	101	101	101	1	45	60
7	[illegible]	NaN	NaN	NaN	NaN	0	NaN	NaN
8	47	47	47	47	47	1	46	46
9	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
10	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
11	402	402	402	402	402	1	36	36
12	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
13	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
14	19	19	19	19	19	1	15	15
15	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
16	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN
17	NaN	NaN	NaN	NaN	NaN	0	NaN	NaN

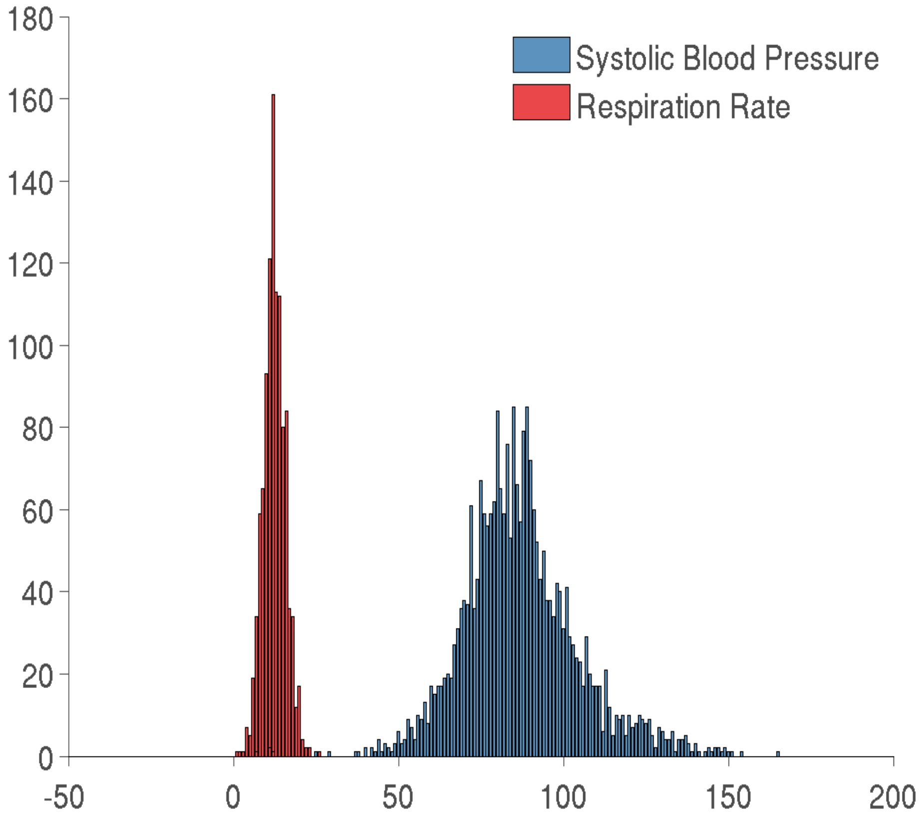
Data come in many forms

- Common pitfalls in medical data:
  - Missing values flagged using a number (-99, 0, 99999),
  - Incorrect units (lb instead of kg),
  - Order of magnitude errors (735 pH instead of 7.35),
  - Sensor artefact (variable dependent).



Cleaning how to?





Handling missing values: removing

- Ignore the feature
  - Pro: Simple, typically not biased
  - Con: May be a very useful feature
- Ignore the example
  - Pro: Simple, all features are kept
  - Con: Removed samples may be biased
  - Con: Data may become small

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
~~Hyundai Sonata 4~~	~~9999~~	~~Korea~~	~~NA~~	~~23~~	~~Medium~~	~~2885~~	~~143~~	~~110~~
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
~~Nissan Maxima V6~~	~~17899~~	~~Japan~~	~~5~~	~~22~~	~~NA~~	~~3200~~	~~180~~	~~160~~
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
~~Oldsmobile Cutlass Supreme V6~~	~~14495~~	~~NA~~	~~1~~	~~21~~	~~Medium~~	~~3220~~	~~189~~	~~135~~
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
~~Chevrolet Lumina APV V6~~	~~13995~~	~~USA~~	~~NA~~	~~18~~	~~Van~~	~~3195~~	~~151~~	~~110~~
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

Introduction to Data Science, Zeev Waks, Intel, Class 1: Data understanding and pre-processing. March 15, 2017.

Handling missing values: imputation

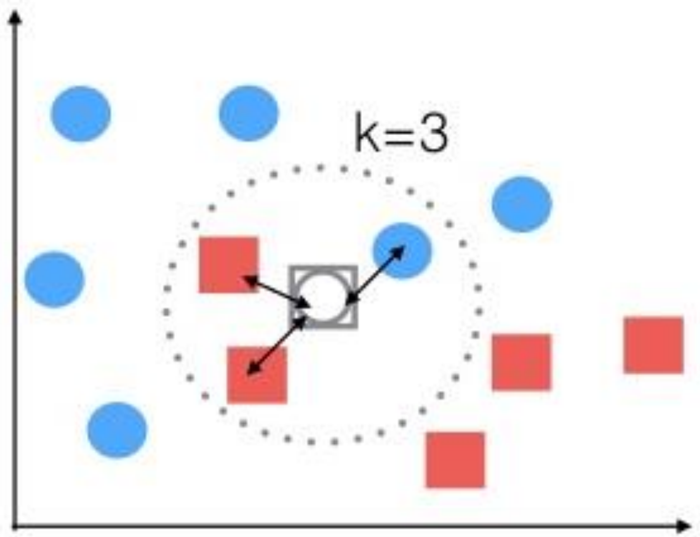
- Estimate the missing values.
- Simple data imputation: **mean, median, mode.**

	Price	Country	Reliability	Mileage	Type	Weight	Disp.	HP
Hyundai Sonata 4	9999	Korea	NA	23	Medium	2885	143	110
Mazda 929 V6	23300	Japan	5	21	Medium	3480	180	158
Nissan Maxima V6	17899	Japan	5	22	NA	3200	180	160
Oldsmobile Cutlass Ciera 4	13150	USA	2	21	Medium	2765	151	110
Oldsmobile Cutlass Supreme V6	14495	NA	1	21	Medium	3220	189	135
Toyota Cressida 6	21498	Japan	3	23	Medium	3480	180	190
Buick Le Sabre V6	16145	USA	3	23	Large	3325	231	165
Chevrolet Caprice V8	14525	USA	1	18	Large	3855	305	170
Ford LTD Crown Victoria V8	17257	USA	3	20	Large	3850	302	150
Chevrolet Lumina APV V6	13995	USA	NA	18	Van	3195	151	110
Dodge Grand Caravan V6	15395	USA	3	18	Van	3735	202	150

Mean (Reliability): (5+5+2+1+3+3+1+3+3)/9 = 2.88
Median (Reliability): 1 1 2 3 3 3 3 5 5
Mode (Country): USA = 6, Japan = 3, Korea = 1.

Handling missing values: K-nearest neighbors

- A similarity based, clustering based approach,
- Distance metric required,
- Fill in missing value using the (median/mean/mode) of the K-nearest neighbors,
- Con: Affected by curse of dimensionality.



Introduction to Data Science, Zeev Waks, Intel, Class 1: Data understanding and pre-processing. March 15, 2017.

Handling missing values

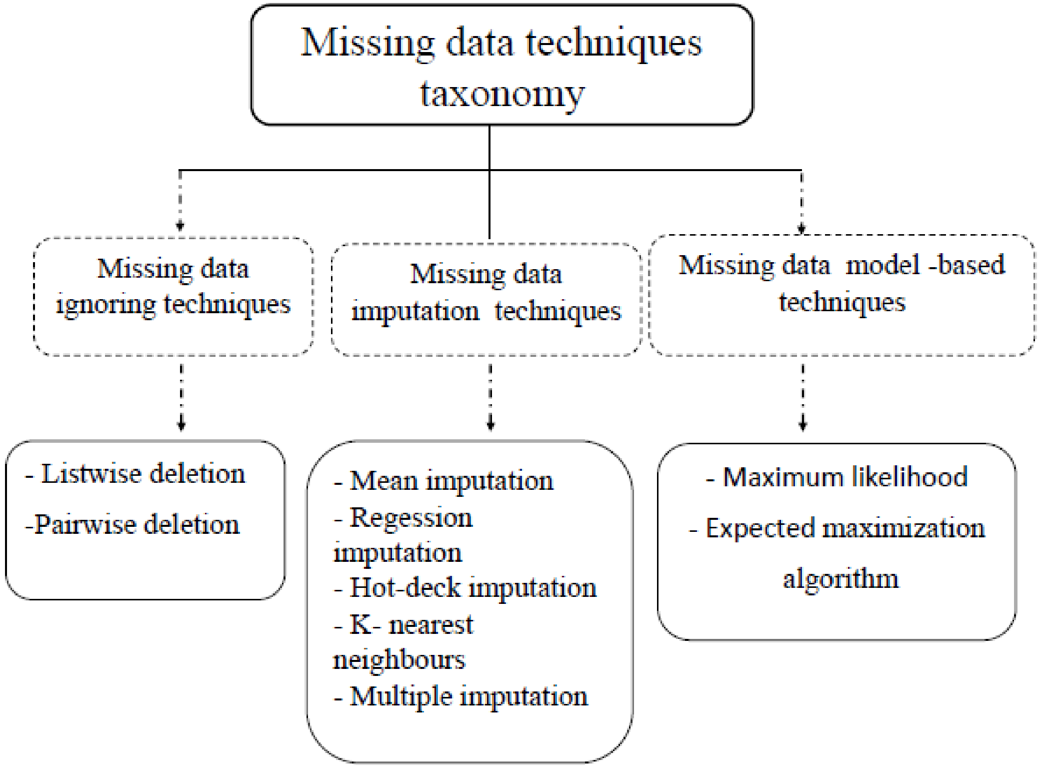


Fig. 1. Missing data techniques taxonomy.

Hooray, Rima, et al. "Handling missing data problems with sampling methods." 2014 International Conference on Advanced Networking Distributed Systems and Applications. IEEE, 2014.

Handling missing values

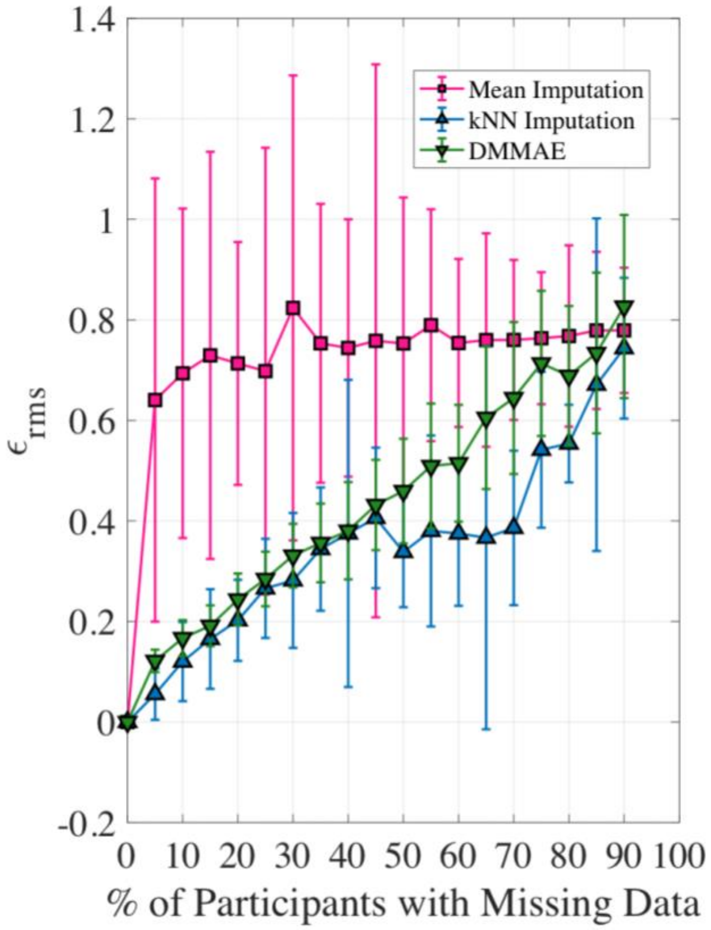


Fig. 2. The errors of missing values resulting from the imputation and DMMAE techniques. These results correspond to the methods outlined in Section 2.5.

- Looking at the approximation of the estimated features in term of the RMSE with respect to the true features.

Prince, J., Andreotti, F., & De Vos, M. (2019, April). Evaluation of Source-wise Missing Data Techniques for the Prediction of Parkinson's Disease Using Smartphones. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3927-3930). IEEE.

Handling missing values

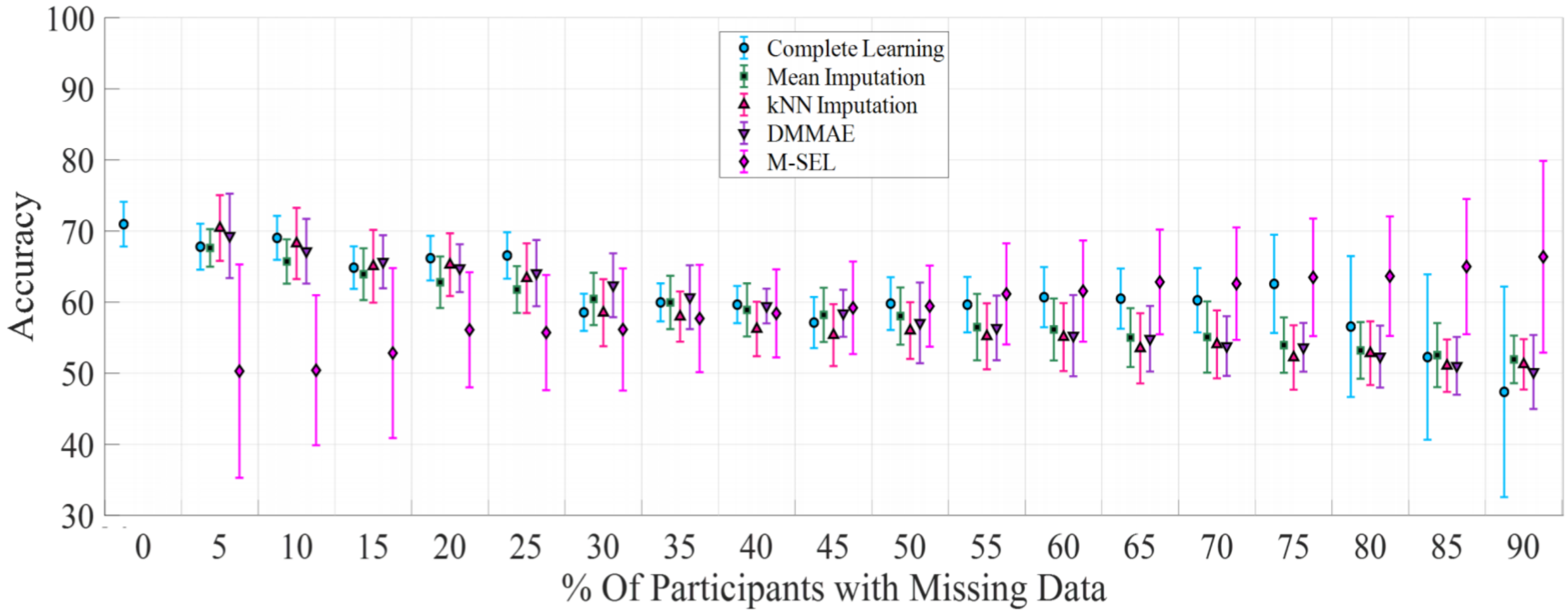
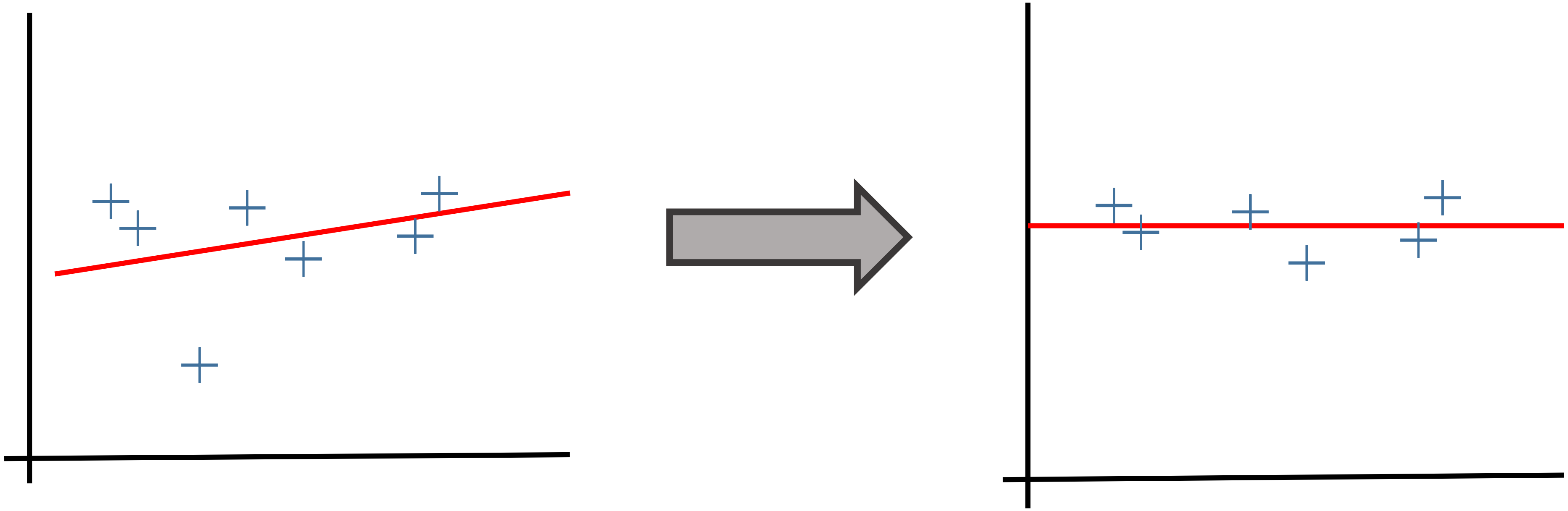


Fig. 1. The binary classification accuracy of each missing data technique at increasing levels of missingness. These results correspond to the methods outlined in Section 2.4.

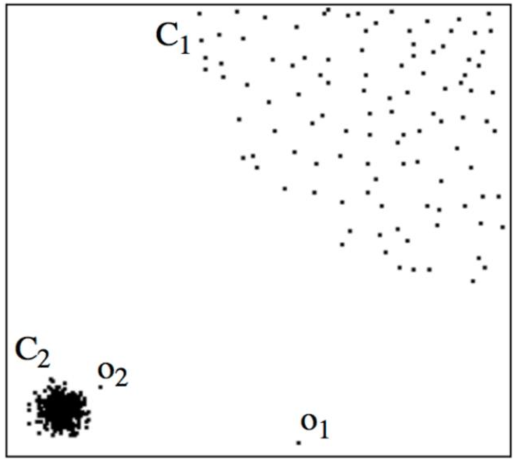
Prince, J., Andreotti, F., & De Vos, M. (2019, April). Evaluation of Source-wise Missing Data Techniques for the Prediction of Parkinson's Disease Using Smartphones. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3927-3930). IEEE.

Detecting outliers

- How do we detect them?
  - Distance based: Look for observation point which are distant from other observations.
  - Build a model (e.g. regression) and look for the observation that are distant from the model. Remove observation with largest error. Repeat.



Detecting outliers

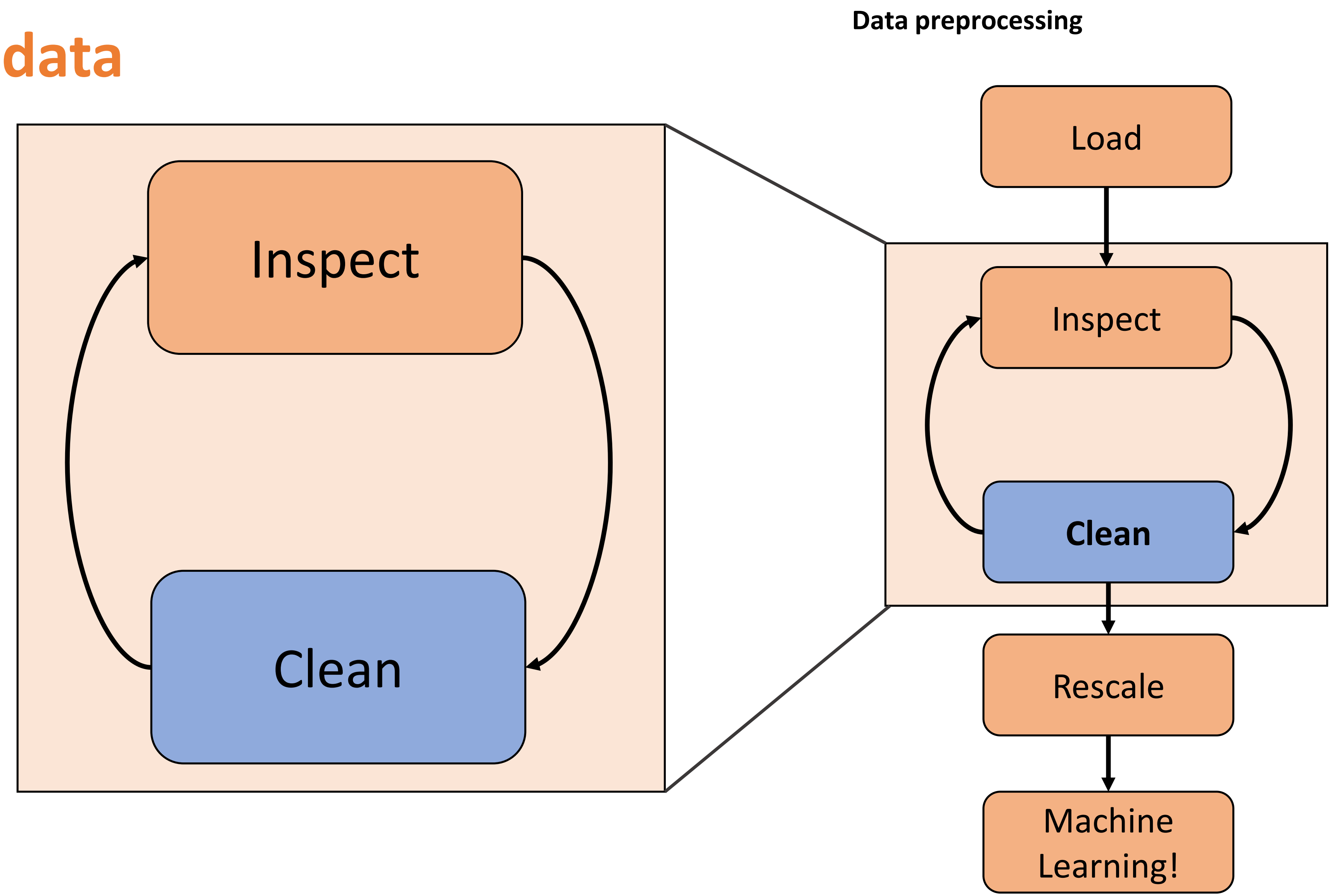


- Also called anomaly detection. Used heavily in fraud, security.
- Examples
  - Density,
  - Data points for which there are fewer than p neighbors within a distance D,
  - Distance
    - The top n data points whose distance to the kth nearest neighbor are the greatest,
    - The top n data points whose average distance to the k nearest neighbors are the greatest.
  - Local Outlier Factor (LOF)
    - Compare the local density of an object to the local densities of its neighbors.

Handling outliers

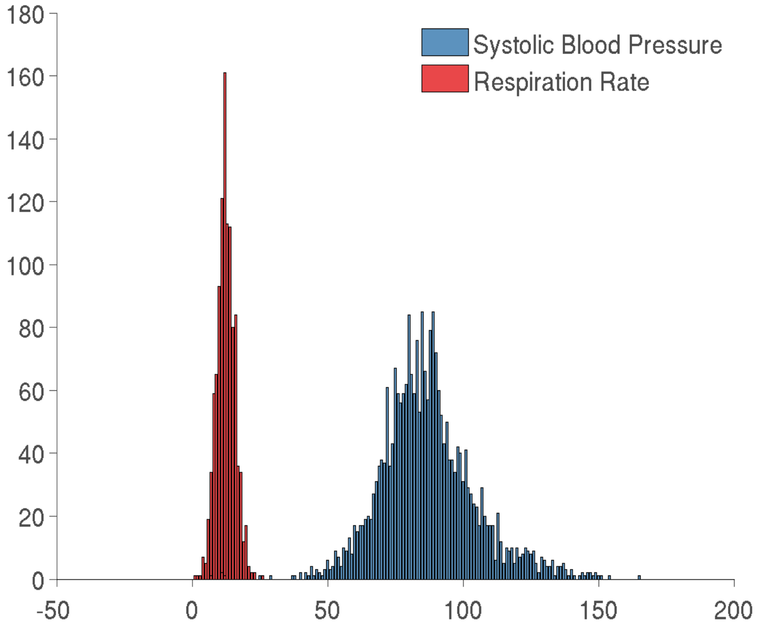
- What to do with them?
  - Remove them: if they are likely errors,
  - Re-weight them: if they are true values but so that they do not affect the model too much.
- Identifying outliers is important for both:
  - Data understanding,
  - Preprocessing.
- Outlier/anomaly detection is a major field of research.

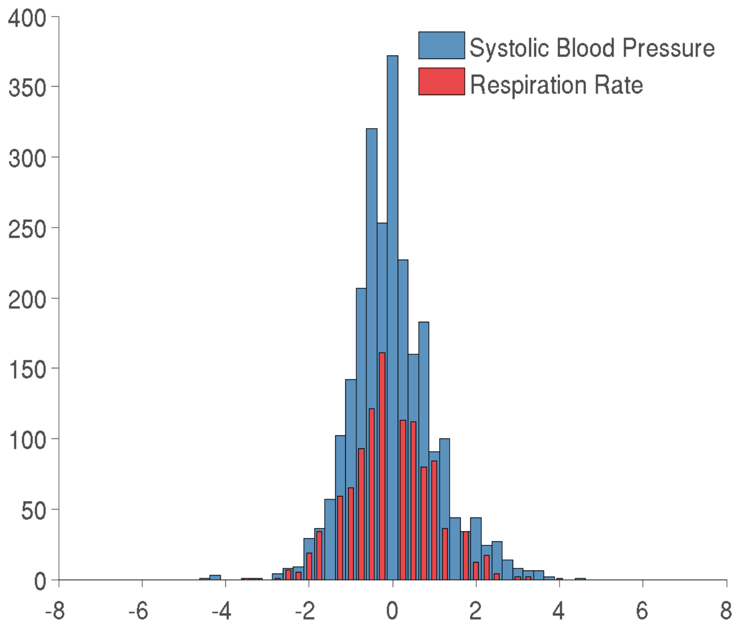
Cleaning data



Rescale

Example: Feature scaling





Is this standardization or normalization?

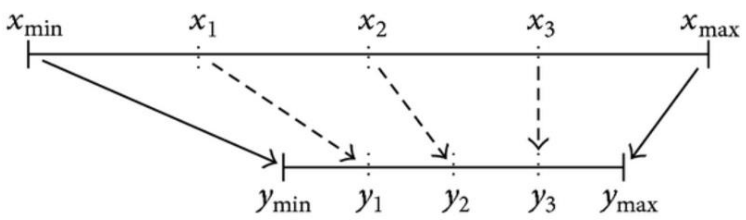
AIMLab.

Feature Scaling: Standardization

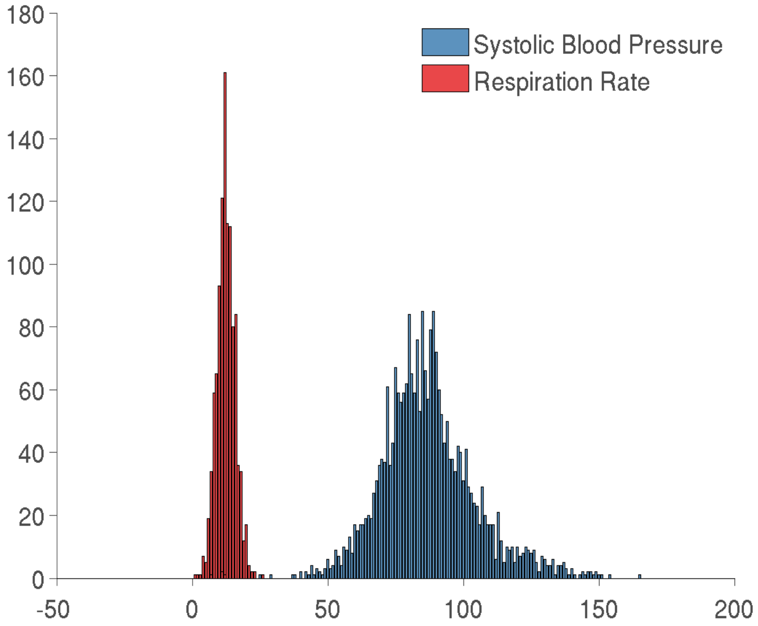
- **Standardization** (or **z-score normalization**): features are rescaled so that they have the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$.
  - $z = \frac{x - \mu}{\sigma}$
- Important for comparing measurements that have different units and it is an assumption for many ML algorithms.
  - Example: gradient descent (used for optimization in LR, SVM and NN), certain weights may update faster than others since the feature values plays a role in the weight updates.
  - $\Delta w_j = -\eta \frac{\partial J}{\partial w_j} = \eta \cdot \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)}) \cdot x_j^{(i)}$

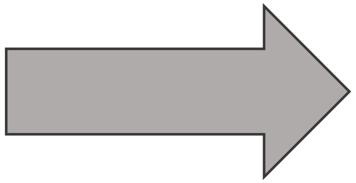
Feature Scaling: Normalization

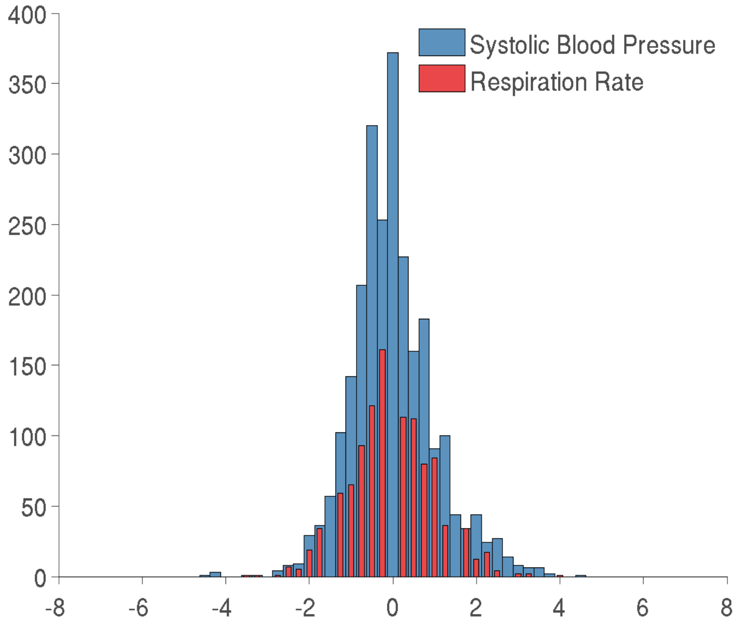
- **Normalization** (or **Min-Max scaling**): rescales the features to [0 1].
  - $x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Useful when there are no outliers with extremely large or small values. For example if the finite set of option for the given feature is {1,2,3,4,5}.
- In some situation we might prefer to map data to a range [-1 1] with zero-mean then we should use the **mean normalization**.
  - $x_{norm} = \frac{x - mean(x)}{\max(x) - \min(x)}$



Example: Feature scaling







Is this standardization or normalization?

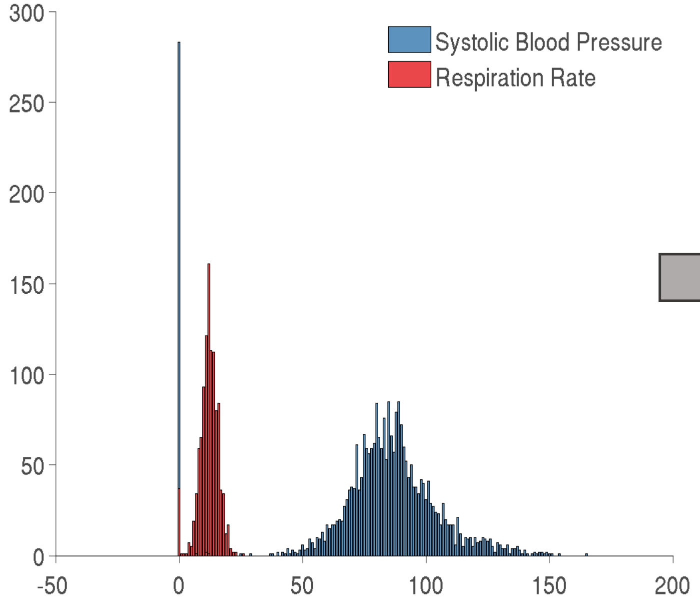
Feature Scaling: Standardization or Normalization?

- Normalization is sensitive to outliers so if there are outliers in the dataset it is not a good idea.
- Standardized data are not bounded (unlike normalization.)
- In practice you often need to experiment!
- Note: there are many other methods for scaling your data. See here:

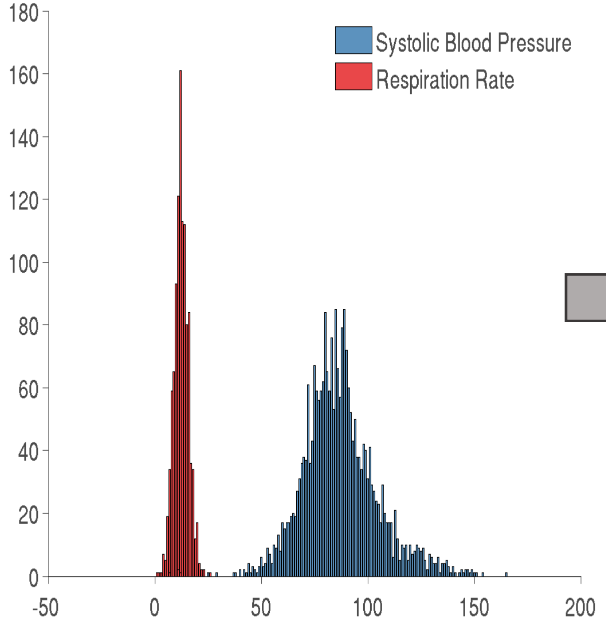
https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py

Going through the pipeline

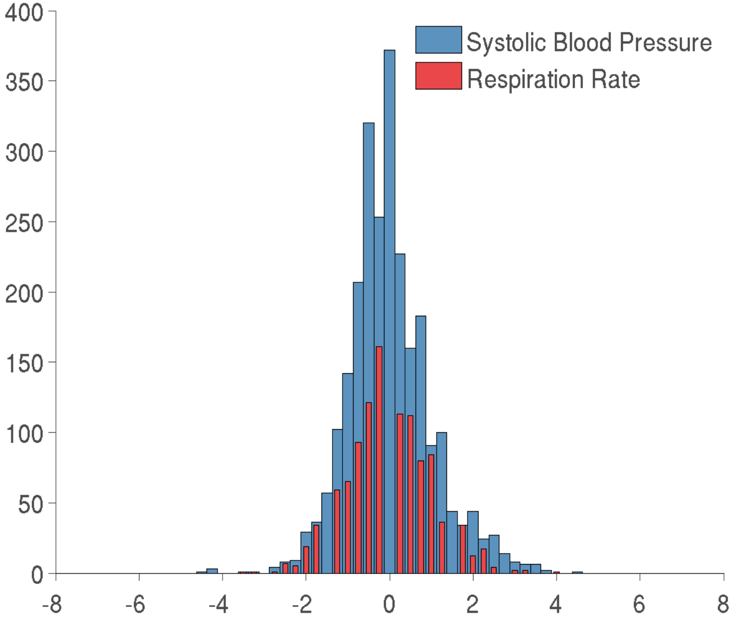
Raw



Clean



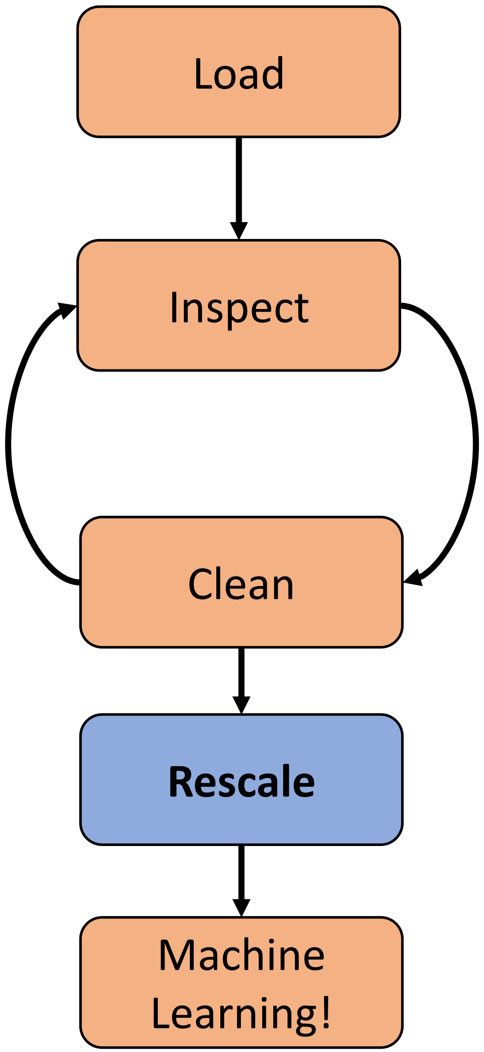
Rescaled



You know

- How to **characterize, represent, inspect, clean and normalize data**.
- How to report **summary statistics and figures** describing your dataset.
- This first step of data exploration and preprocessing is crucial to any ML project.

Data preprocessing



Take home

- Data come in different types. Clarify which one.
- Describe your data using summary statistics and data visualization tools.
    - Five Number Summary in statistics: min, Q1, median, Q3, max. Add a 6th number: the number of examples (m).
    - Standard visualization tools: Line Plot, Bar Chart, Histogram Plot, Boxplot and Scatter Plot.
    - For numerical data, histograms are the best way to look at your data, boxplots are the second best way. Barplots is the worst way. Don’t use them! Violin plot is another good option.
- Use summary statistics and visualization tools to understand your data and flag any abnormality.

AIMLab.

- Look for abnormalities: **missing values, outliers and incorrect entries**.
  - Missing values: removing, imputation, K-nearest neighbor etc.
  - Outliers: anomaly detection.
- Never use data you don't understand!
- After cleaning inspect again and recomputed your updated summary statistics.
- **Standardization** and **normalization** have the same aim: to build features that have similar ranges.
- In statistics, standardization is the subtraction of the mean and then dividing by its standard deviation. In algebra, normalization is the process of dividing of the vector by its length and it transforms your data in the range 0-1.
- Both have drawbacks: **normalization is sensitive to outliers** so if there are outliers in the dataset it is not a good idea. **Standardized data are not bounded** (unlike normalization.)

References

[1] Introduction to Data Science, Zeev Waks, Intel, Class 1: Data understanding and pre-processing. March 15, 2017.

[2] https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

[3] Oxford, CDT course 2015.

[4] Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python. Jason Brownlee, 2018.