



Machine Learning in Healthcare


## #C02 Data Exploration and Preprocessing

Technion-IIT, Haifa, Israel

---

Assist. Prof. Joachim Behar  
Biomedical Engineering Faculty  
Technion-IIT

AIMLab  




---

---

---

---


---

---

---

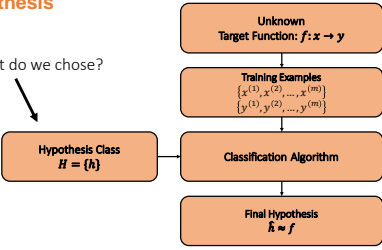
---

Data exploration and preprocessing

AIMLab 

### Hypothesis

What do we chose?



```

graph TD
    A[Unknown Target Function: f: x -> y] --> B["Training Examples  
{x^{(1)}, x^{(2)}, ..., x^{(m)}  
y^{(1)}, y^{(2)}, ..., y^{(m)}}"]
    C[Hypothesis Class  
H = {h}] --> D[Classification Algorithm]
    B --> D
    D --> E[Final Hypothesis  
h approx f]
  
```

2

---

---

---

---


---

---

---

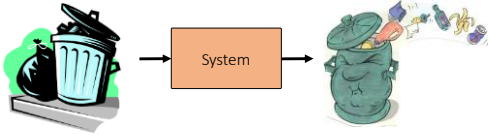
---

Data exploration and preprocessing

AIMLab 

### The importance of preprocessing

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



```

graph LR
    A[Garbage in] --> B[System]
    B --> C[Garbage out]
  
```

3

---

---

---

---

---

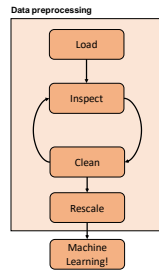
---

---

---

## Topics covered

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



4

---

---

---

---

---

---

---

## Data types

5

---

---

---

---

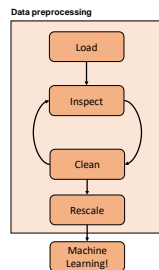
---

---

---

## Topics covered

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



6

---

---

---

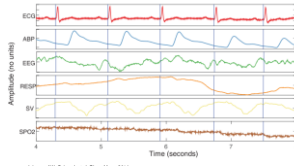
---

---

---

---

## Sources of data



Johnson AM, Sellar J et al. Phys. Meas. 2014



Index	1	2	3	4	5	6	7	8	9
1	100	100	100	100	100	100	100	100	100
2	100	100	100	100	100	100	100	100	100
3	100	100	100	100	100	100	100	100	100
4	100	100	100	100	100	100	100	100	100
5	100	100	100	100	100	100	100	100	100
6	100	100	100	100	100	100	100	100	100
7	100	100	100	100	100	100	100	100	100
8	100	100	100	100	100	100	100	100	100
9	100	100	100	100	100	100	100	100	100
10	100	100	100	100	100	100	100	100	100
11	100	100	100	100	100	100	100	100	100
12	100	100	100	100	100	100	100	100	100
13	100	100	100	100	100	100	100	100	100
14	100	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	100	100	100
17	100	100	100	100	100	100	100	100	100
18	100	100	100	100	100	100	100	100	100
19	100	100	100	100	100	100	100	100	100
20	100	100	100	100	100	100	100	100	100
21	100	100	100	100	100	100	100	100	100
22	100	100	100	100	100	100	100	100	100
23	100	100	100	100	100	100	100	100	100
24	100	100	100	100	100	100	100	100	100
25	100	100	100	100	100	100	100	100	100
26	100	100	100	100	100	100	100	100	100
27	100	100	100	100	100	100	100	100	100
28	100	100	100	100	100	100	100	100	100
29	100	100	100	100	100	100	100	100	100
30	100	100	100	100	100	100	100	100	100

Physician ICU dataset

7

---

---

---

---

---

---

---

---

---

---

## 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

8

---

---

---

---

---

---

---

---

---

---

## Exploratory data analysis

9

---

---

---

---

---

---

---

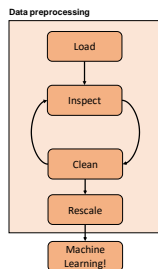
---

---

---

## Topics covered

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



10

---

---

---

---

---

---

---

---

## Exploratory data analysis

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

11

---

---

---

---

---

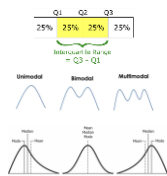
---

---

---

## Summary statistics

- **Central tendency:** mean, median and mode.
- Measure of **variability/spread**:
  - Range: min, max.
  - Variance and standard deviation.
  - Interquartile range (IQR). Used to measure the spread or variability within your data.
- Modality: number of peaks it contains.
- Skewness: measure of the symmetry of a distribution.
- Kurtosis: measures whether your dataset is heavy-tailed or light-tailed compared to a normal distribution.
- **Typically you will report the "Five Number Summary statistics": min, Q1, median, Q3, max.**



12

---

---

---

---

---

---

---

---

## 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.

13

---

---

---

---

---

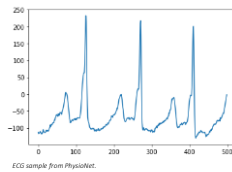
---

---

---

## 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.



14

---

---

---

---

---

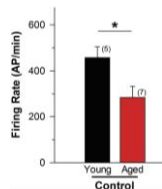
---

---

---

## Bar chart

- Present relative quantities for multiple categories.
- The x-axis represents the categories and the y-axis represents the quantity for each category.



15

---

---

---

---

---

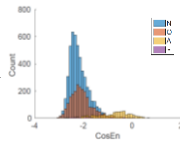
---

---

---

## 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. Crit C 2017

Histograms are density estimates. A density estimate gives a good impression of the distribution of the data [...]. The idea is to locally represent the data density by counting the number of observations in a sequence of consecutive intervals (bins) ...

Page 11, Applied Multivariate Statistical Analysis, 2015.

16

---

---

---

---

---

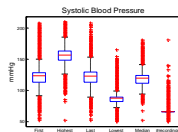
---

---

---

## Boxplots

- Summarize the distribution of a data sample.
- The x-axis is used to represent the data sample. The y-axis represents the observed values.
- 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. A line is also drawn within the box at the 50th 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.
- Boxplot is a graphical representation of the Five Number Summary.**



17

---

---

---

---

---

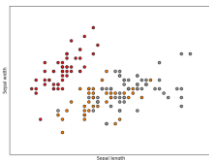
---

---

---

## Scatter Plot

- Used to summarize the relationship between two paired data samples.
- 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 observation.



18

---

---

---

---

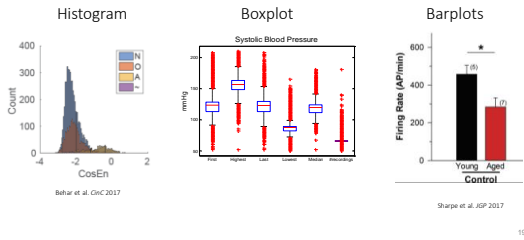
---

---

---

---

## Histogram, boxplot or barplot for my paper?




---

---

---

---

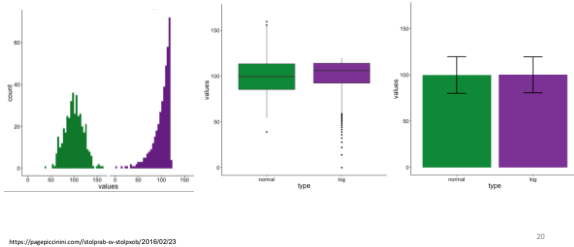
---

---

---

---

## Note: Histogram, boxplot or barplot for my paper?




---

---

---

---

---

---

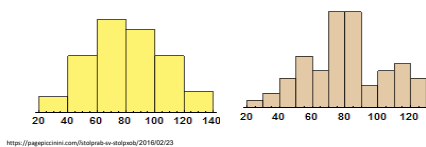
---

---

## 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




---

---

---

---

---

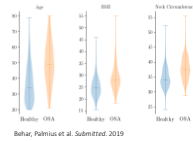
---

---

---

## Violin plots

- Similar to boxplots except that they also 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.



22

---

---

---

---

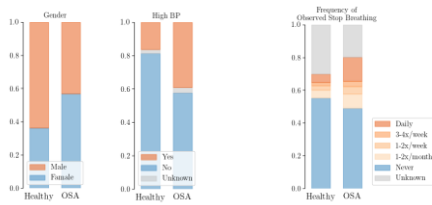
---

---

---

---

## Binary and ordinal features



23

---

---

---

---

---

---

---

---

Handle abnormalities

24

---

---

---

---

---

---

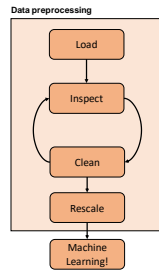
---

---



## Topics covered

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



25

---

---

---

---

---

---

---

---

## Clean

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



26

---

---

---

---

---

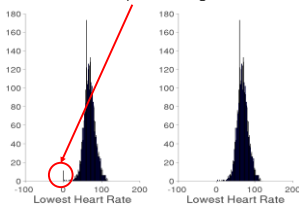
---

---

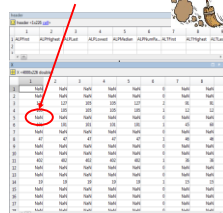
---

## Look for the dust!

Incorrect entry or missing value



Missing values



27

---

---

---

---

---

---

---

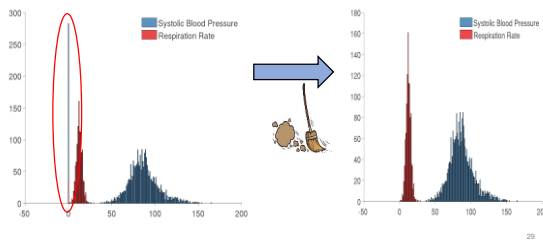
---

## 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).

28

## Cleaning how to?



29

## Handling missing values: removing

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

	Price	Country	Make	Model	Year	Type	Weight	Displ	HP
<del>toyota corolla</del>	23999	Japan	Toyota	Corolla	2010	Small	2800	180	158
<del>toyota prius</del>	19999	Japan	Toyota	Prius	2010	Small	2400	158	158
<del>toyota camry</del>	24999	Japan	Toyota	Camry	2010	Medium	2500	180	158
<del>toyota sienna</del>	24999	Japan	Toyota	Sienna	2010	Large	3500	232	158
<del>toyota tundra</del>	24999	Japan	Toyota	Tundra	2010	Large	5600	401	300
<del>toyota tacoma</del>	24999	Japan	Toyota	Tacoma	2010	Medium	3400	232	158
<del>toyota 4runner</del>	24999	Japan	Toyota	4Runner	2010	Medium	3400	232	158
<del>toyota sequoia</del>	24999	Japan	Toyota	Sequoia	2010	Large	5600	401	300
<del>toyota yaris</del>	19999	Japan	Toyota	Yaris	2010	Small	1800	158	158
<del>toyota prius v</del>	19999	Japan	Toyota	Prius V	2010	Small	1800	158	158
<del>toyota prius x</del>	19999	Japan	Toyota	Prius X	2010	Small	1800	158	158
<del>toyota prius s</del>	19999	Japan	Toyota	Prius S	2010	Small	1800	158	158
<del>toyota prius l</del>	19999	Japan	Toyota	Prius L	2010	Small	1800	158	158
<del>toyota prius e</del>	19999	Japan	Toyota	Prius E	2010	Small	1800	158	158
<del>toyota prius h</del>	19999	Japan	Toyota	Prius H	2010	Small	1800	158	158
<del>toyota prius i</del>	19999	Japan	Toyota	Prius I	2010	Small	1800	158	158
<del>toyota prius j</del>	19999	Japan	Toyota	Prius J	2010	Small	1800	158	158
<del>toyota prius k</del>	19999	Japan	Toyota	Prius K	2010	Small	1800	158	158
<del>toyota prius l</del>	19999	Japan	Toyota	Prius L	2010	Small	1800	158	158
<del>toyota prius m</del>	19999	Japan	Toyota	Prius M	2010	Small	1800	158	158
<del>toyota prius n</del>	19999	Japan	Toyota	Prius N	2010	Small	1800	158	158
<del>toyota prius o</del>	19999	Japan	Toyota	Prius O	2010	Small	1800	158	158
<del>toyota prius p</del>	19999	Japan	Toyota	Prius P	2010	Small	1800	158	158
<del>toyota prius q</del>	19999	Japan	Toyota	Prius Q	2010	Small	1800	158	158
<del>toyota prius r</del>	19999	Japan	Toyota	Prius R	2010	Small	1800	158	158
<del>toyota prius s</del>	19999	Japan	Toyota	Prius S	2010	Small	1800	158	158
<del>toyota prius t</del>	19999	Japan	Toyota	Prius T	2010	Small	1800	158	158
<del>toyota prius u</del>	19999	Japan	Toyota	Prius U	2010	Small	1800	158	158
<del>toyota prius v</del>	19999	Japan	Toyota	Prius V	2010	Small	1800	158	158
<del>toyota prius w</del>	19999	Japan	Toyota	Prius W	2010	Small	1800	158	158
<del>toyota prius x</del>	19999	Japan	Toyota	Prius X	2010	Small	1800	158	158
<del>toyota prius y</del>	19999	Japan	Toyota	Prius Y	2010	Small	1800	158	158
<del>toyota prius z</del>	19999	Japan	Toyota	Prius Z	2010	Small	1800	158	158

30

## Handling missing values: imputation

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

	Price	Country	Reliability	mileage	Type	weight	Displ.	mpg
hyundai sonata 4	9890	korea	na	22	medium	1885	143	110
honda 924 v6	12300	japan	1	22	medium	1880	180	134
nissan maxima v6	17890	japan	1	22	na	2300	180	180
oldsmobile cutlass ciera 4	13150	usa	2	22	medium	2760	133	110
oldsmobile cutlass supreme v6	14450	na	1	22	medium	3220	180	133
toyota corolla 4	24400	japan	3	23	medium	1480	180	190
honda le sdn v6	16140	usa	2	19	large	1925	210	169
chevrolet caprice v8	14320	usa	1	18	large	1815	300	170
ford LTD Crown Victoria v8	17220	usa	1	20	large	4050	300	120
chevrolet lumina apv v6	13995	usa	na	18	vgn	3195	131	110
dodge crane caravan v6	13395	usa	3	18	vgn	2735	200	130

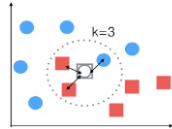
Mean (Reliability):  $(5+5+2+1+3+3+1+3+3)/9 = 2.88$

Median (Reliability): 1 1 2 3 **3** 3 3 5 5

Mode (Country): **USA=6**, Japan = 3, Korea = 1.

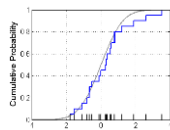
## Handling missing values: K-nearest neighbors

- 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



## Handling missing values: sampling

- Find the cumulative distribution function  $F(b)$  (remember,  $0 = F(-\infty) \leq F(b) \leq F(\infty) = 1$ )
- Generate random number from a uniform distribution  $[0,1]$
- Calculate  $b = F^{-1}(c)$



Works for both discrete and continuous distributions

## Handling missing values

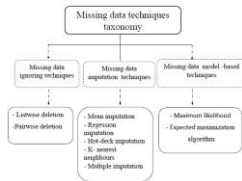


Fig. 1. Missing data techniques taxonomy.

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

34

---

---

---

---

---

---

---

---

---

---

## Handling missing values

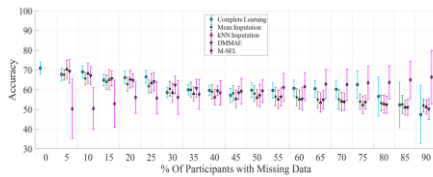


Fig. 1. The binary classification accuracy of each missing data technique at increasing levels of missingness. These results correspond to the methods outlined in Section 2.4.

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.

35

---

---

---

---

---

---

---

---

---

---

## Handling missing values

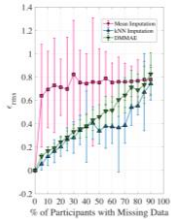


Fig. 2. The errors of missing values resulting from the imputation and DMMAE techniques. These results correspond to the methods outlined in Section 2.5.

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.

36

---

---

---

---

---

---

---

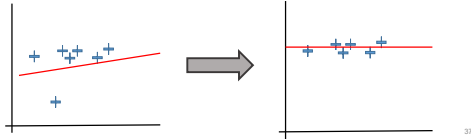
---

---

---

## Handling 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



37

---

---

---

---

---

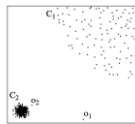
---

---

---

## Handling outliers

- Also sometimes 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  $k$ th 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.



38

---

---

---

---

---

---

---

---

## 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,
  - Pre-processing.
- Outlier/anomaly detection is a major field of research.

39

---

---

---

---

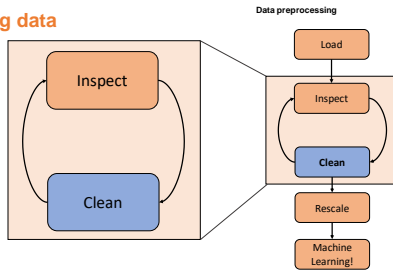
---

---

---

---

## Cleaning data



40

---

---

---

---

---

---

---

---

## Rescale

41

---

---

---

---

---

---

---

---

## 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 (most?) 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.
- $\Delta w_j = -\eta \frac{\partial J}{\partial w_j} = \eta \sum_i (y^{(i)} - \hat{y}^{(i)}) \cdot x_j^{(i)}$

42

---

---

---

---

---

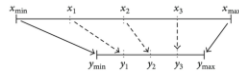
---

---

---

## 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 options for the given feature is {1,2,3,4,5}.
  - $x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$
- In some situation we might prefer to map data to a range [-1 1] with zero-mean then we should use the mean normalization.



43

---

---

---

---

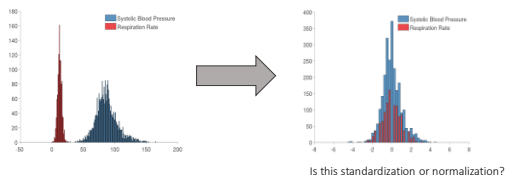
---

---

---

---

## Example: Feature Scaling



44

---

---

---

---

---

---

---

---

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

45

---

---

---

---

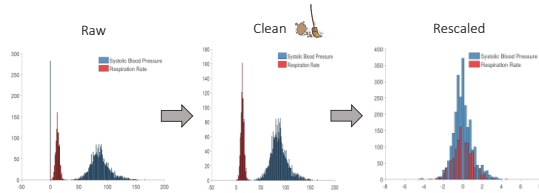
---

---

---

---

## Going through the pipeline



46

---

---

---

---

---

---

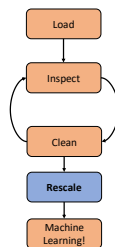
---

---

## 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



47

---

---

---

---

---

---

---

---

## 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.

48

---

---

---

---

---

---

---

---



## Take home

- 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 recompute 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.)

49

---

---

---

---

---

---

---

---

## 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](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.

50

---

---

---

---

---

---

---

---