Data Wangling

Dr. Christan Grant
Dr. Laura Melissa Cruz Castro
CAP5771 – Introduction to Data Science
University of Florida

Outline

- Overview of Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Discretization

Data Quality: Why Preprocess the Data?

Measures for Data Quality: A Multidimensional View

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, …
- Consistency: some modified but some not, dangling, …
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

Data cleaning

Detect and correct missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data reduction

Dimensionality reduction	Numerosity reduction	Data compression

Data transformation and data discretization

Normalization	Concept hierarchy generation

Outline

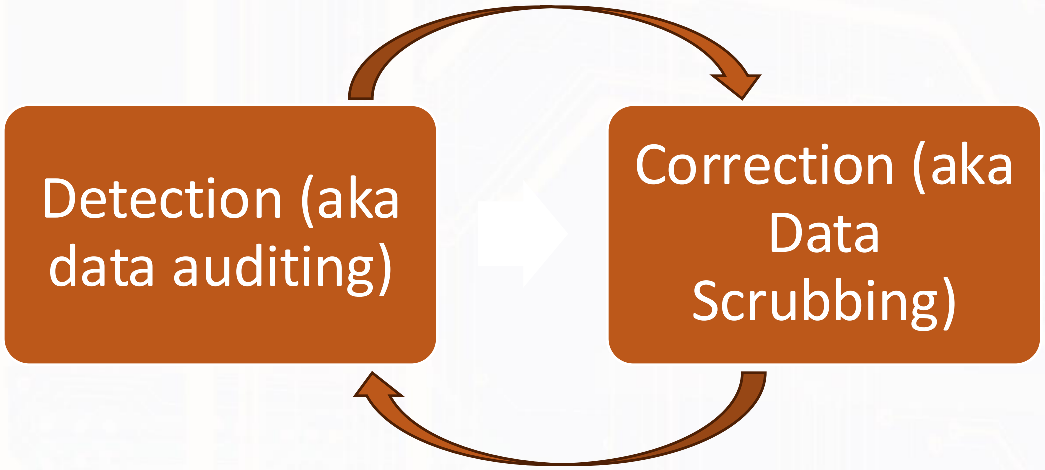
- Overview of Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Discretization

Data Quality Issues Examples

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

- **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., Occupation=“ ” (missing data)
- **Noisy**: containing noise, errors, or outliers
  - e.g., Salary=“−10” (an error)
- **Inconsistent**: containing discrepancies in codes or names, e.g.,
  - Age=“42”, Birthday=“03/07/2010”
  - Was rating “1, 2, 3”, now rating “A, B, C”
  - Discrepancy between duplicate records
- **Intentional**: (e.g., disguised missing data)
  - Jan. 1 as everyone’s birthday?

Data Cleaning as a Process



Data discrepancy detection

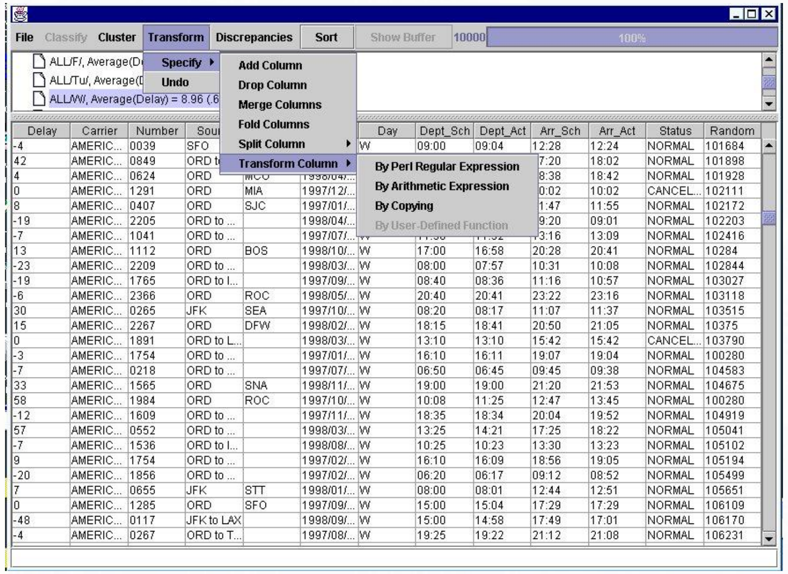
- Use metadata (e.g., domain, range, dependency, distribution)
- Check constraints and rules on data (e.g., functional dependency constraints, uniqueness rule)
- Outlier detection through correlation/distribution/clustering analysis

Data correction

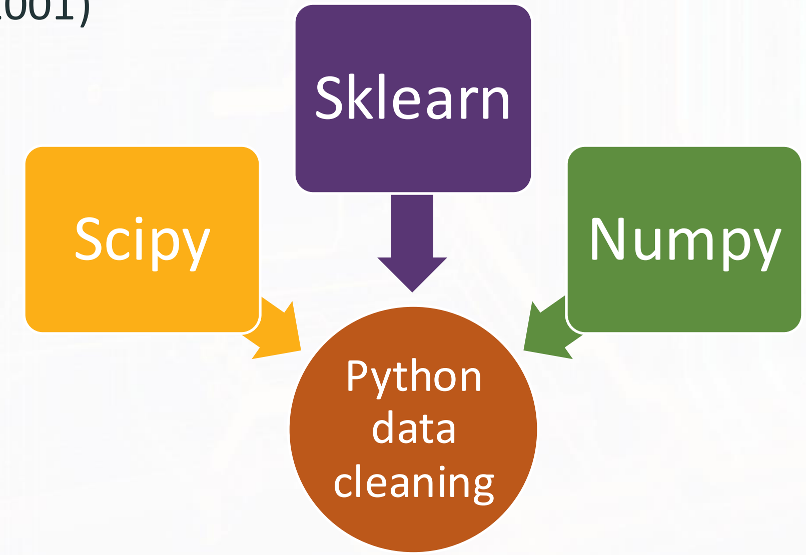
- Binning, regression, clustering
- Human-in-the-loop inspection and correction

Data Cleaning is an iterative process

- **Continuous Improvement**: Ongoing activity that enhances data quality and usability.
- **Adaptive Procedures**: Adjusts methods to address emerging data issues.
- **Feedback Integration**: Refines cleaning strategies based on analysis outcomes.



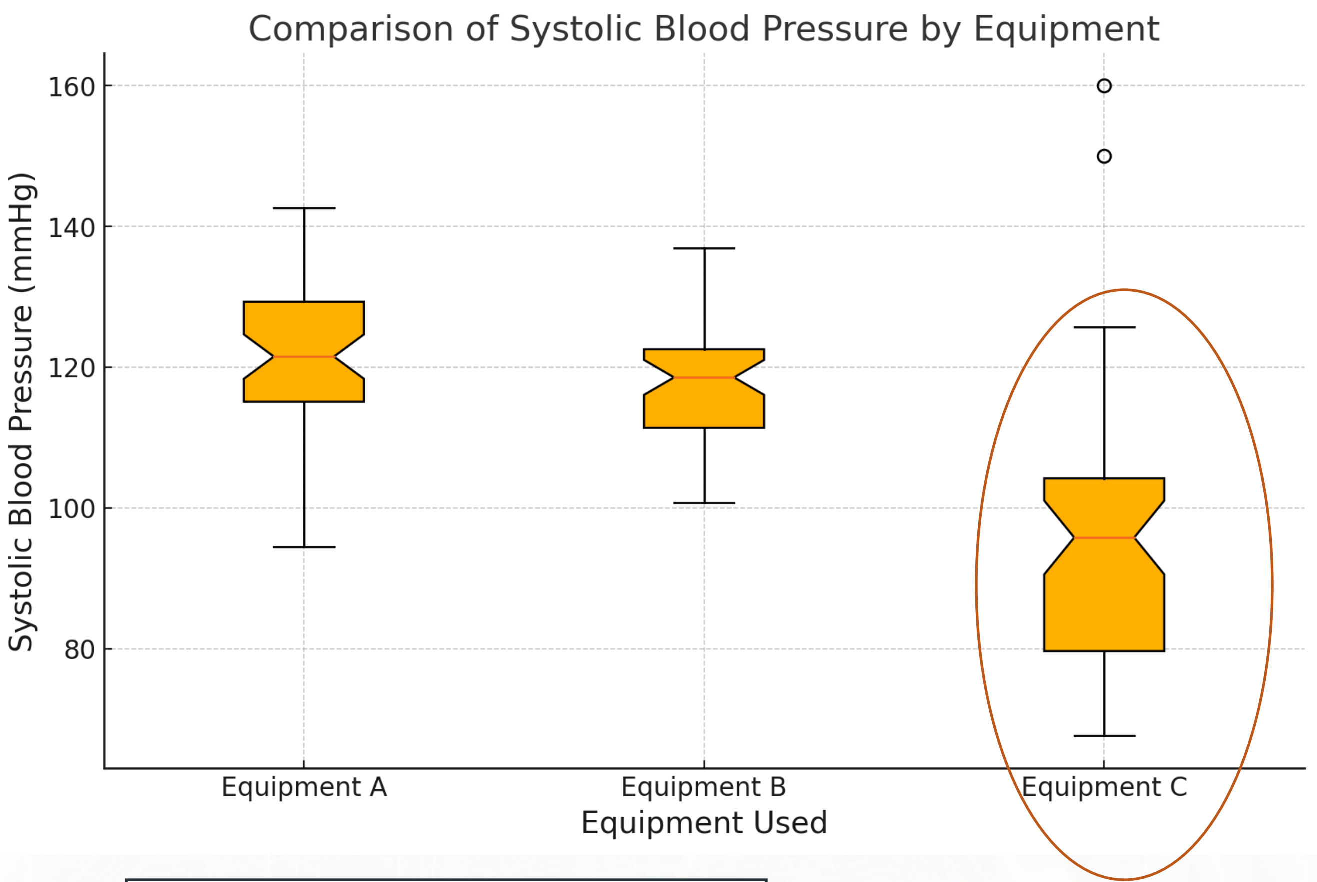
Potter's wheel GUI (2001)



Python libraries for data cleaning

Data discrepancy detection using metadata

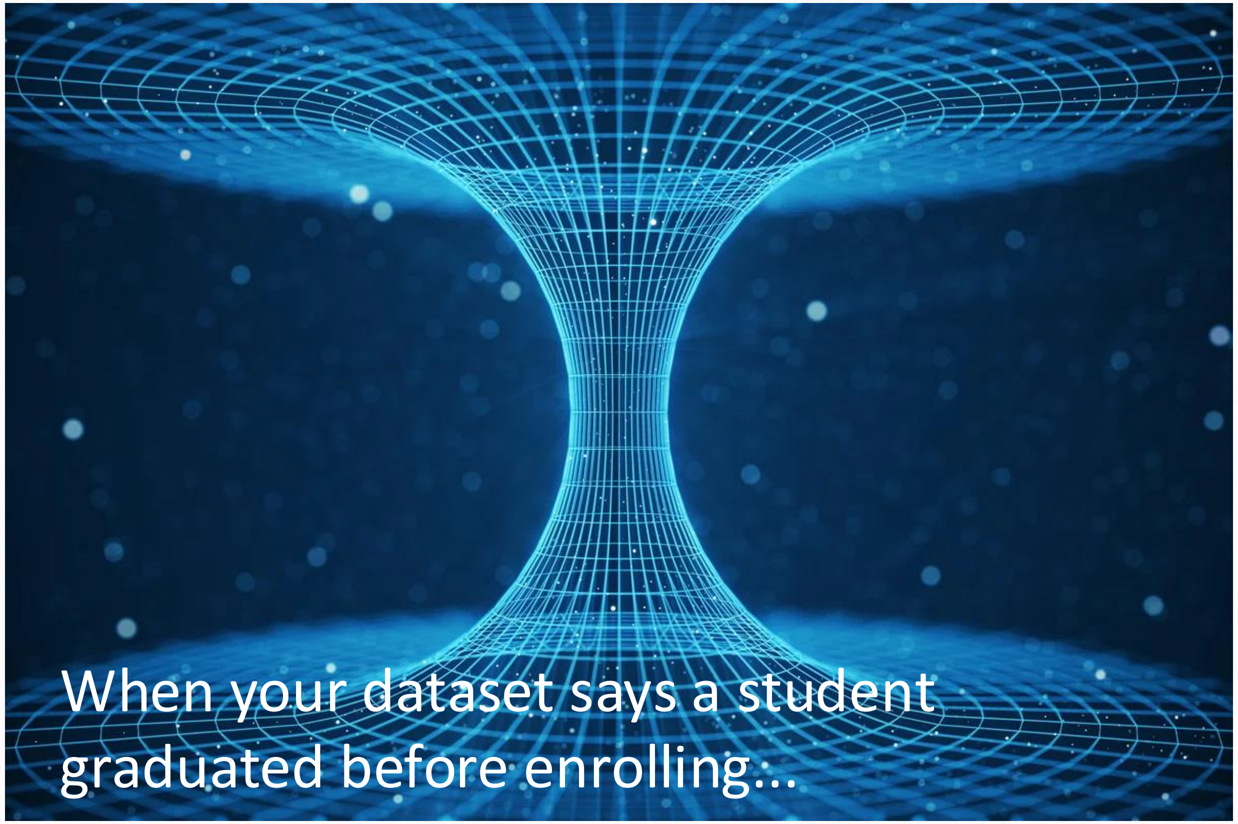
In a medical research data Python libraries like **pydicom** are used to extract metadata from **DICOM** files, including scan dates and equipment details. These elements are then compared to a dataset containing patient test results to identify mismatches in scan dates or equipment used, ensuring the integrity and accuracy of medical research data by aligning actual conditions with recorded data.



We would not know that this discrepancy existed due to equipment if we did not have access to the metadata!

Data discrepancy detection using rules

- Knowing that data must be within certain ranges e.g. glucose must be more than 54 mg/dL
- Knowing that the data must follow a rule with respect to another variable e.g. Student graduation must be after student enrollment date
- Knowing that there should not be duplicates e.g. social security number



Outlier detection

Statistical Methods:

- IQR (Interquartile Range): Identify as outliers any data points that lie more than 1.5 times the IQR below the first quartile or above the third quartile.
- Z-Score: Consider data points that have a Z-score (standard deviations from the mean) greater than 3 as outliers.

Visual Methods:

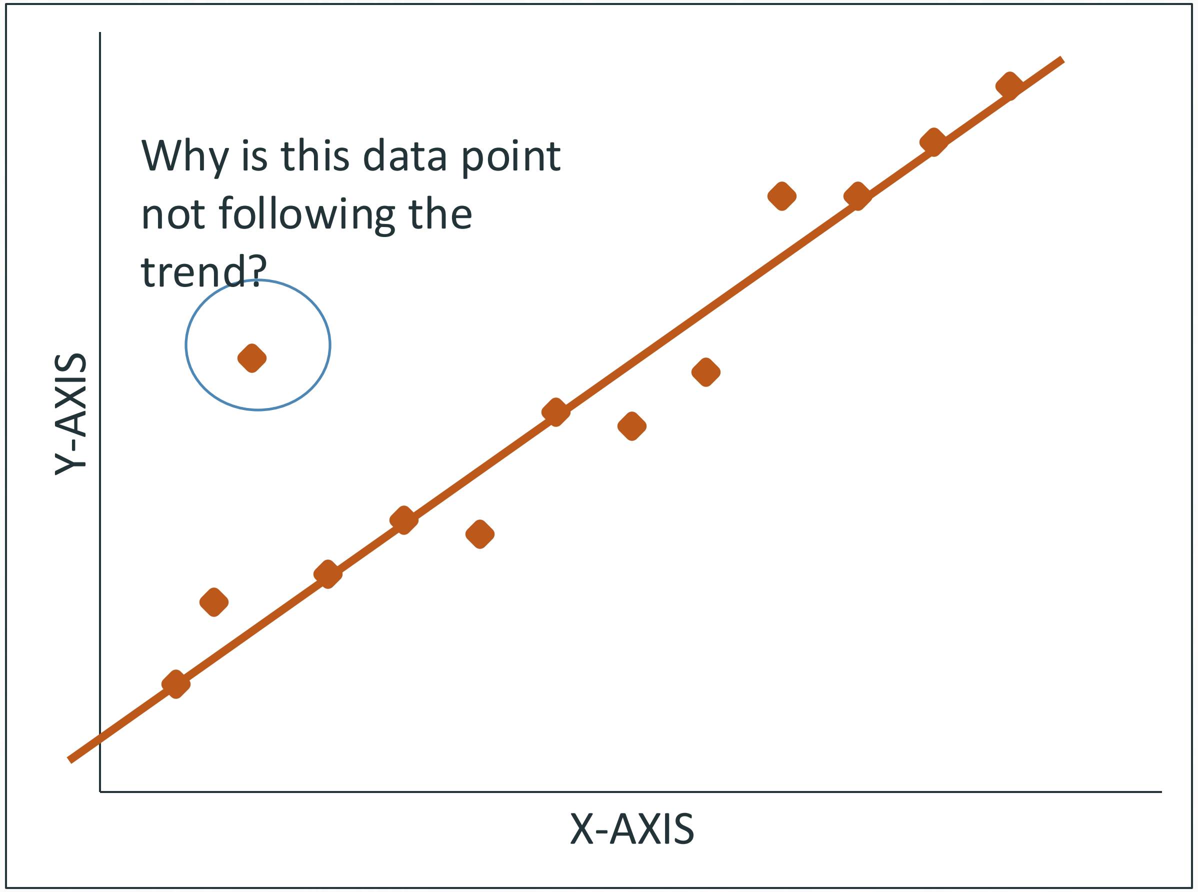
- Box Plots: Use box plots to visually identify data points that lie outside the whiskers, typically 1.5*IQR from the quartiles.
- Scatter Plots: Observe for data points that deviate significantly from the group pattern.

Correlation:

- Outliers in Correlation: Detect single points that can significantly change the correlation coefficient between variables, indicating their potential as outliers.

Clustering Techniques:

- DBSCAN or K-Means: Use clustering algorithms where outliers will not fit well into any cluster or will form very small clusters away from the majority.



How to Correct Dirty Data?

- Binning
- Regression
- Clustering
- Combined computer and human inspection

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into **equal-frequency (equi-depth) bins**:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

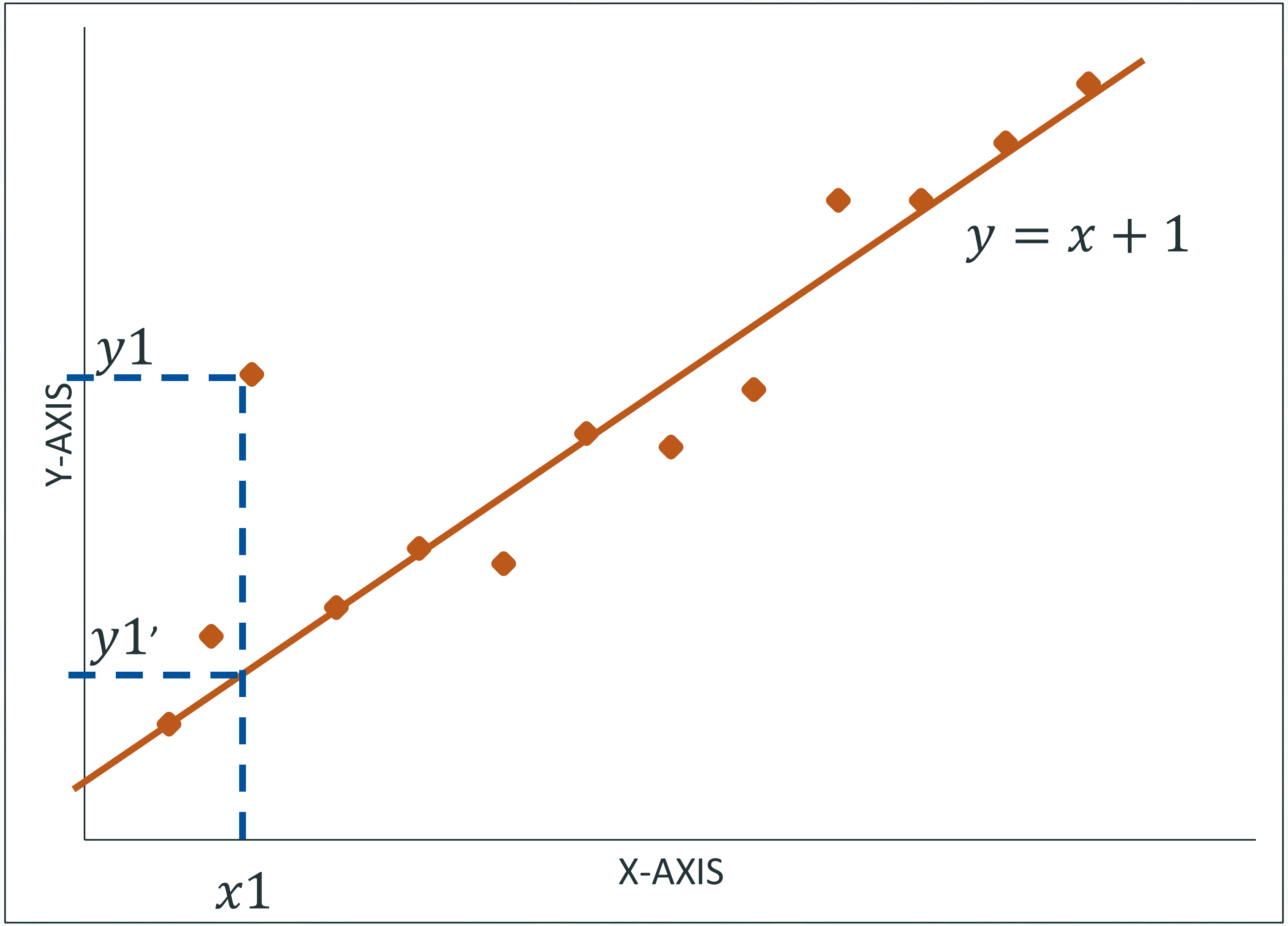
Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

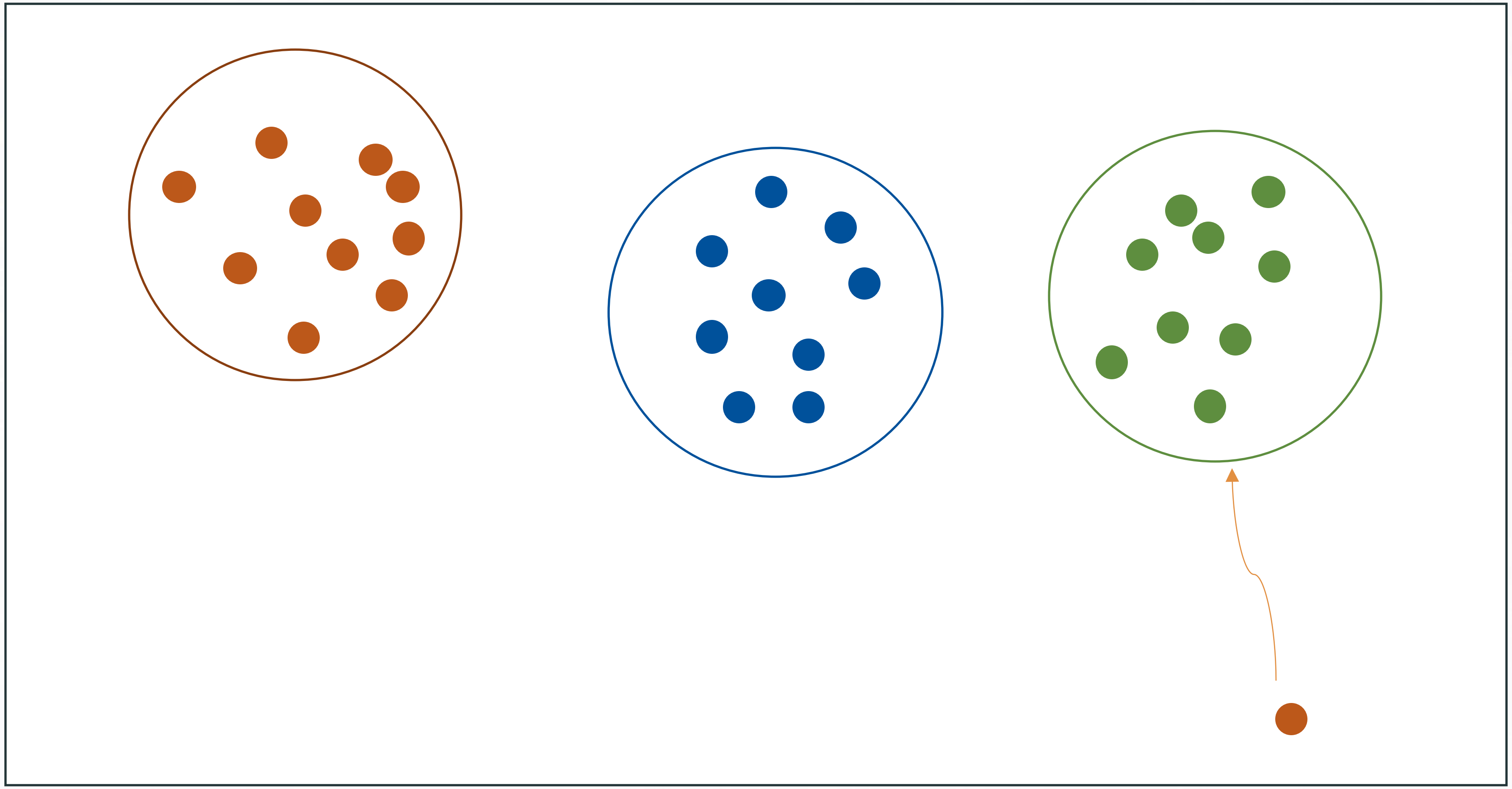
Regression For Data Smoothing



While we show here the easiest case, in which we use a linear regression model, more complex models can be used for data smoothing. However, it is critical to have a **strong hypothesis about the relationship** between the data to be inputted and the auxiliar variable(s) to consider this method of imputation.

Clustering for data smoothing

Raw Data



The mean (or sometimes the median) of the closest cluster is then used to replace or adjust values that are deemed outliers or incorrect within that dataset. This can be particularly **useful in scenarios where data points are expected to form distinct groups,** and deviations from these groups are considered errors or outliers.