



Data Wangling

Dr. Christan Grant

Dr. Laura Melissa Cruz Castro

CAP5771 – Introduction to Data Science

University of Florida



Overview of Data Preprocessing



Data Cleaning



Data Integration



Data Reduction



Data Transformation and Discretization

Measures for Data Quality: A Multidimensional View

Accuracy: correct or wrong, accurate or not

Completeness: not recorded, unavailable, ...

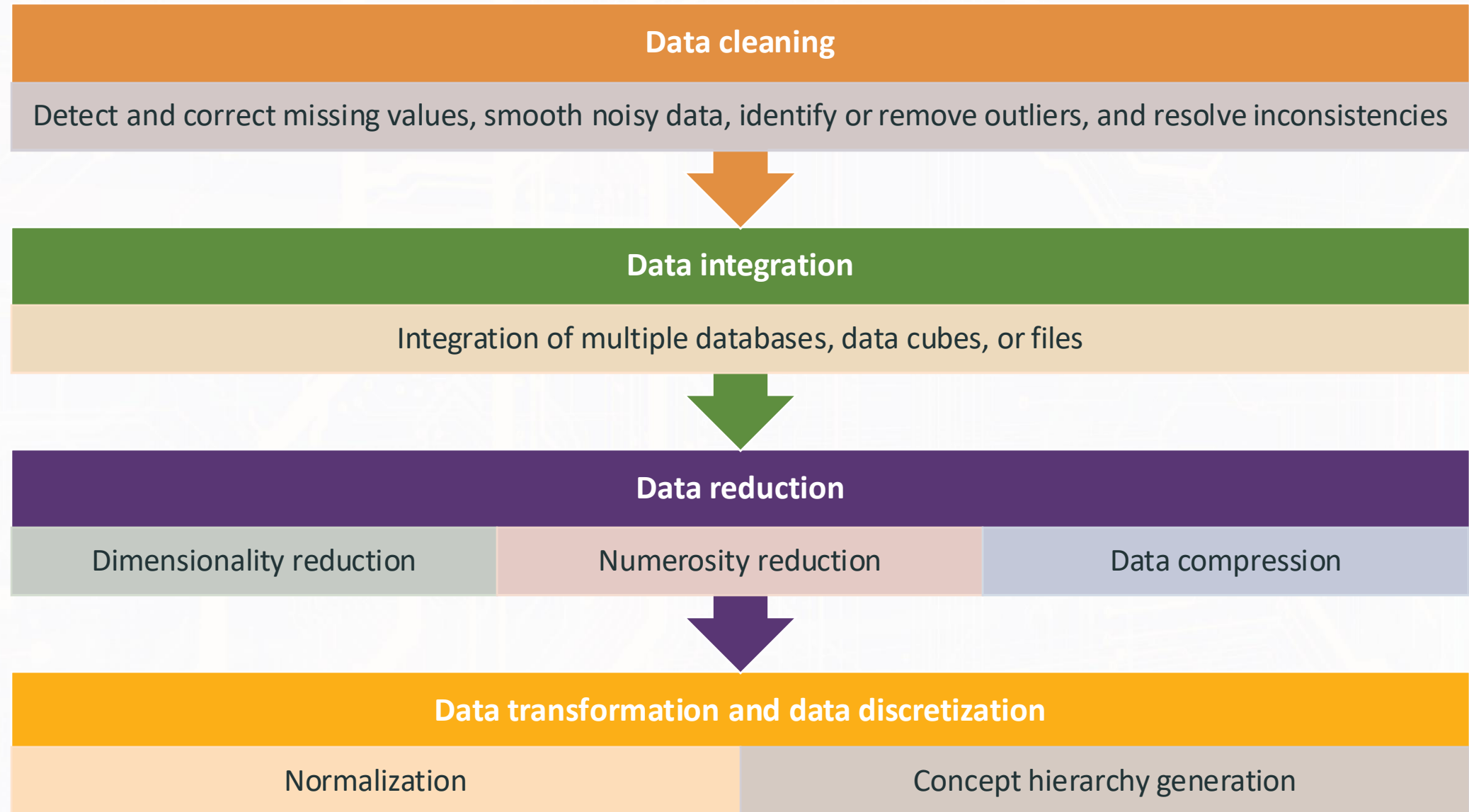
Consistency: some modified but some not, dangling, ...

Timeliness: timely update?

Believability: how trustable the data are correct?

Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing





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"

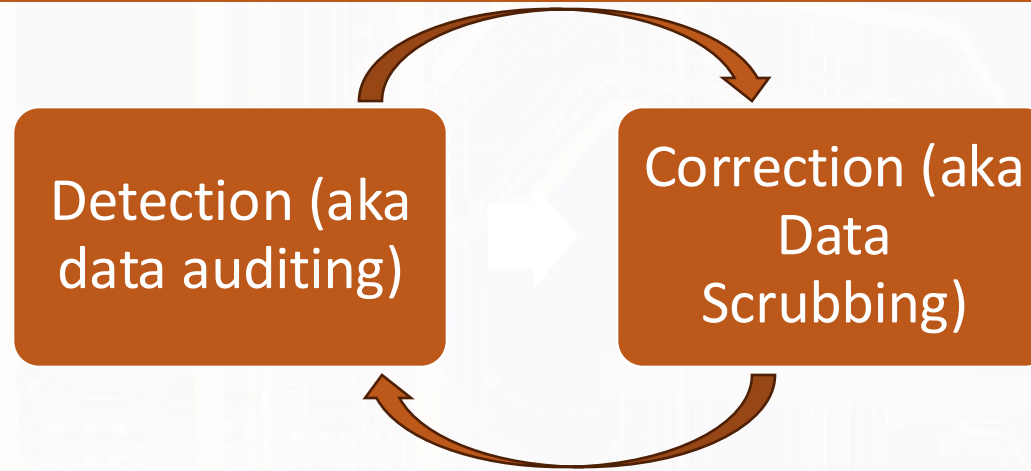
Was rating "1, 2, 3", now rating "A, B, C"

Discrepancy between duplicate records

Intentional: (e.g., disguised missing data)

Jan. 1 as everyone's birthday?

Data Cleaning as a Process



Data discrepancy detection

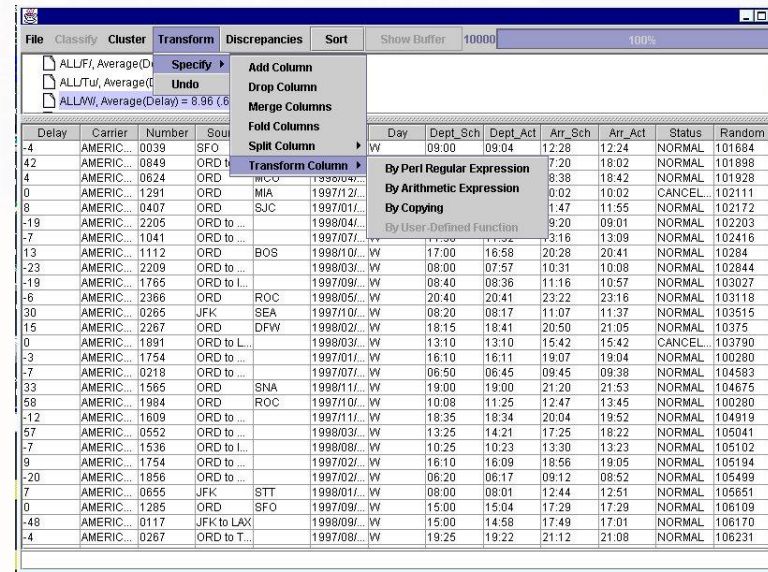
- Use metadata (e.g., domain, range, dependency, distribution)
- Check constraints and rules on data (e.g., functional dependency constraints, uniqueness rule)
- Outlier detection through correlation/distribution/clustering analysis

Data correction

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

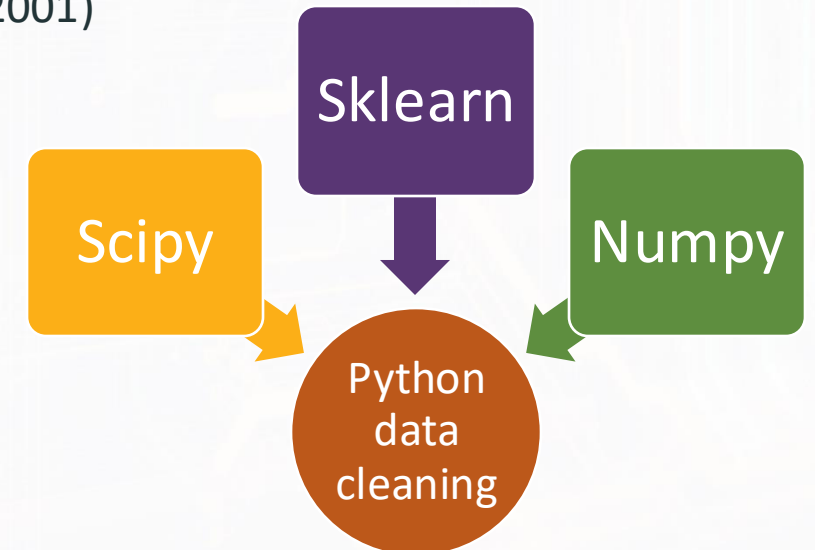
Data Cleaning is an iterative process

- **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.



Delay	Carrier	Number	Source	Day	Dept_Sch	Dept_Act	Arr_Sch	Arr_Act	Status	Random
-4	AMERIC...	0039	SFO	1997/07/...	09:00	09:04	12:28	12:24	NORMAL	101684
42	AMERIC...	0849	ORD to ...	1998/03/...	08:00	07:57	10:31	10:08	NORMAL	101898
4	AMERIC...	0624	ORD to ...	1997/11/...	08:38	08:42	10:42	10:38	NORMAL	101928
0	AMERIC...	1291	ORD to ...	1997/11/...	08:02	08:02	10:02	10:02	CANCEL	102111
8	AMERIC...	0407	ORD to ...	1997/01/...	11:47	11:55	11:55	11:55	NORMAL	102172
-19	AMERIC...	2205	ORD to ...	1998/04/...	09:20	09:01	11:02	10:58	NORMAL	102203
-7	AMERIC...	1041	ORD to ...	1997/07/...	13:16	13:09	13:09	13:09	NORMAL	102416
13	AMERIC...	1112	ORD to ...	1998/10/...	17:00	16:58	20:28	20:41	NORMAL	10284
-23	AMERIC...	2209	ORD to ...	1998/03/...	08:00	07:57	10:31	10:08	NORMAL	102844
-19	AMERIC...	1765	ORD to L...	1997/09/...	08:40	08:36	11:16	10:57	NORMAL	103027
-6	AMERIC...	2366	ORD to ...	1998/05/...	20:40	20:41	23:22	23:16	NORMAL	103118
30	AMERIC...	0265	JFK to ...	1997/10/...	08:20	08:17	11:07	11:37	NORMAL	103515
15	AMERIC...	2267	ORD to ...	1998/02/...	18:15	18:41	20:50	21:05	NORMAL	10375
0	AMERIC...	1891	ORD to L...	1998/03/...	13:10	13:10	15:42	15:42	CANCEL	103790
-3	AMERIC...	1754	ORD to ...	1997/01/...	16:10	16:11	19:07	19:04	NORMAL	100280
-7	AMERIC...	0218	ORD to ...	1997/07/...	06:50	06:45	09:45	09:38	NORMAL	104583
33	AMERIC...	1565	ORD to ...	1998/11/...	19:00	19:00	21:20	21:53	NORMAL	104675
58	AMERIC...	1984	ORD to ...	1997/10/...	11:25	12:47	13:45	13:45	NORMAL	100280
-12	AMERIC...	1609	ORD to ...	1997/11/...	18:35	18:34	20:04	19:52	NORMAL	104819
57	AMERIC...	0552	ORD to ...	1998/03/...	13:25	14:21	17:25	18:22	NORMAL	105041
-7	AMERIC...	1536	ORD to L...	1998/08/...	10:25	10:23	13:30	13:23	NORMAL	105102
9	AMERIC...	1754	ORD to ...	1997/02/...	16:10	16:09	18:56	19:05	NORMAL	105194
-20	AMERIC...	1856	ORD to ...	1997/02/...	06:20	06:17	09:12	08:52	NORMAL	105499
7	AMERIC...	0655	JFK to ...	1998/01/...	08:00	08:01	12:44	12:51	NORMAL	105651
0	AMERIC...	1285	ORD to ...	1997/09/...	15:00	15:04	17:29	17:29	NORMAL	106109
-48	AMERIC...	0117	JFK to LAX	1998/09/...	15:00	14:58	17:49	17:01	NORMAL	106170
-4	AMERIC...	0267	ORD to T...	1997/08/...	19:25	19:22	21:12	21:08	NORMAL	106231

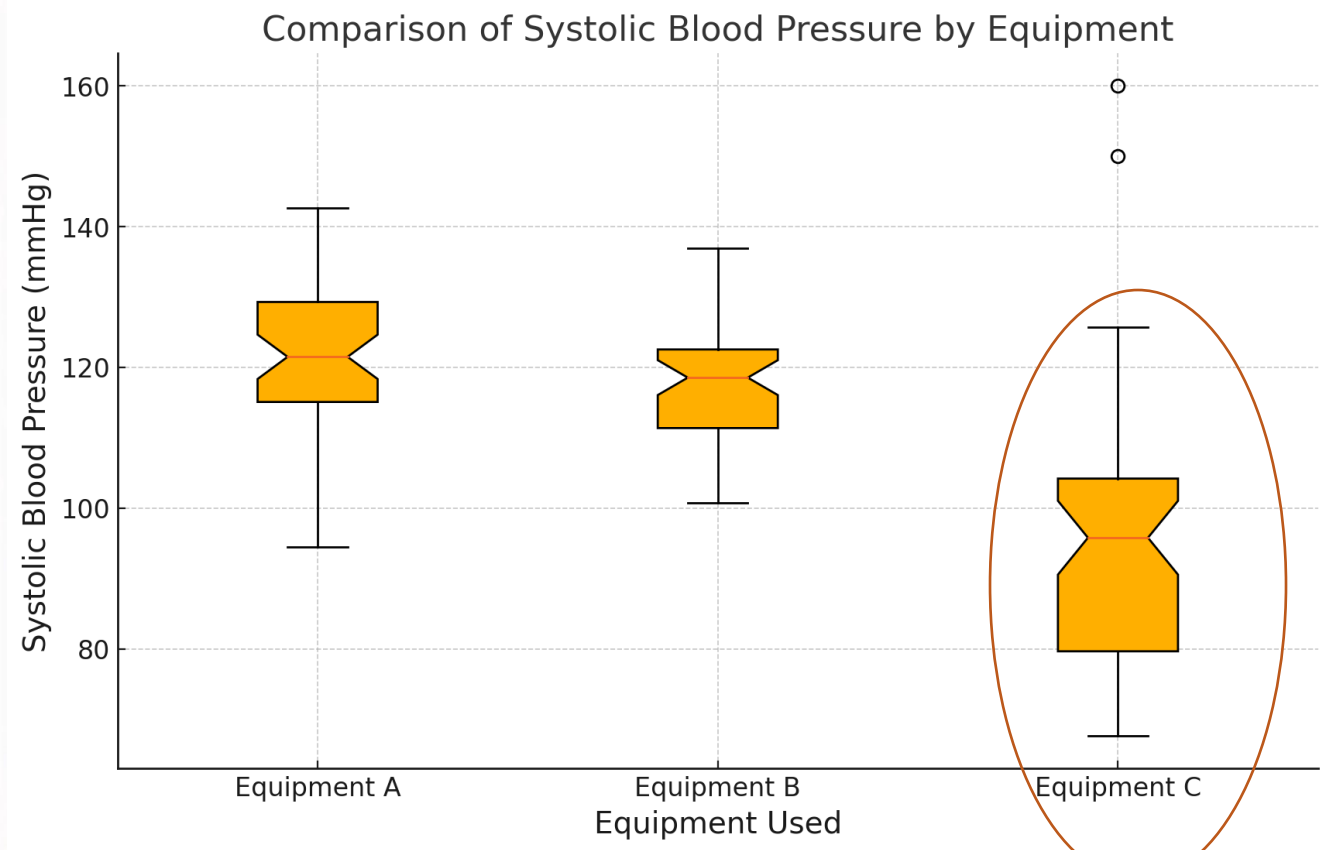
Potter's wheel GUI (2001)



Python libraries for data cleaning

Data discrepancy detection using metadata

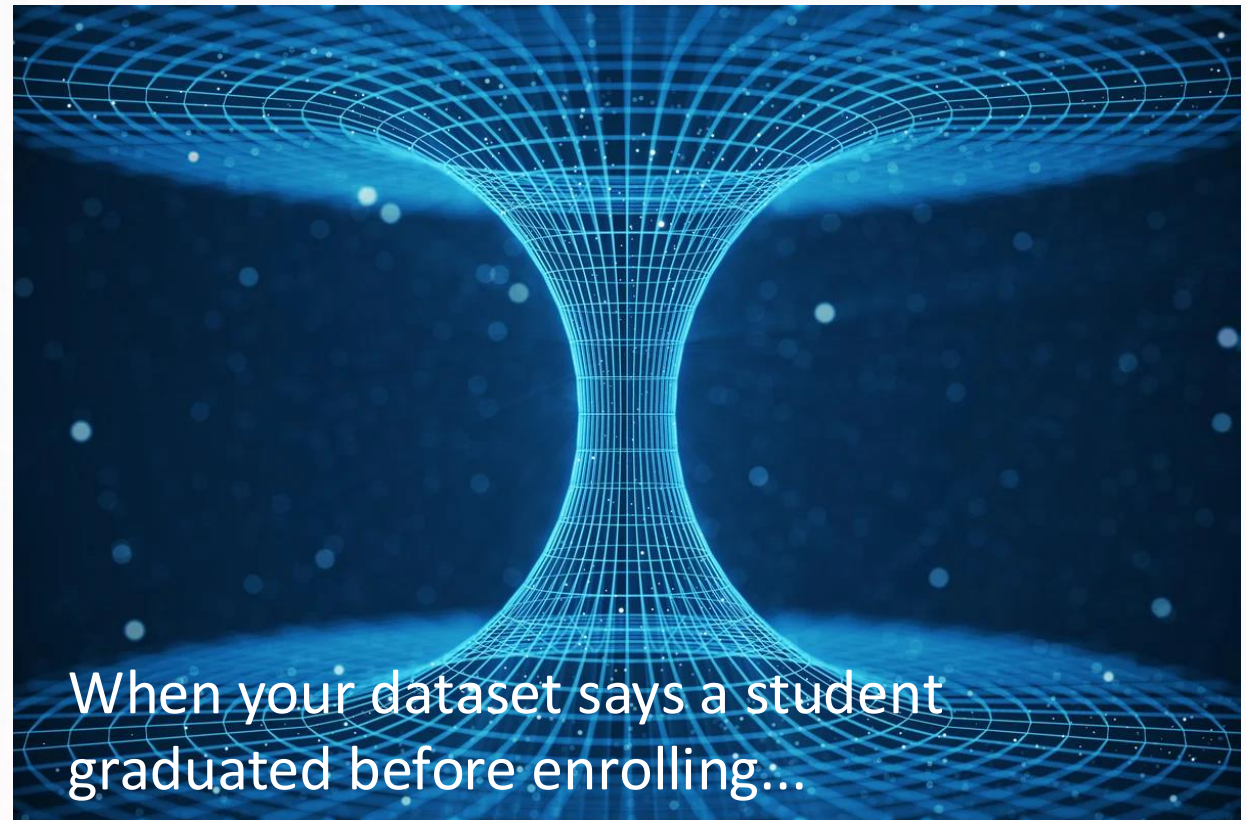
In a medical research data Python libraries like **pydicom** are used to extract metadata from **DICOM** files, including scan dates and equipment details. These elements are **then compared to a dataset containing patient test results to identify mismatches in scan dates or equipment used**, ensuring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



When your dataset says a student graduated before enrolling...

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:

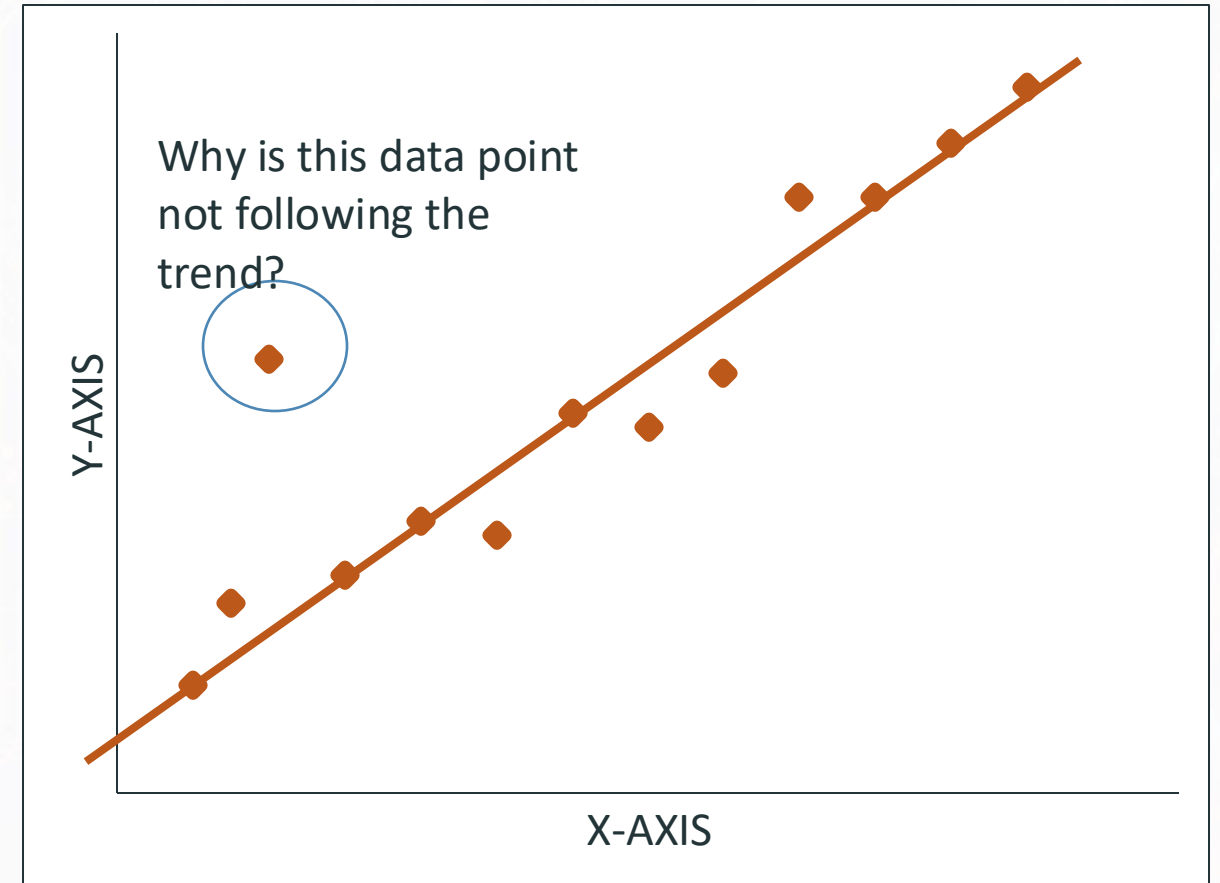
- Box Plots: Use box plots to visually identify data points that lie outside the whiskers, typically $1.5 \times \text{IQR}$ from the quartiles.
- Scatter Plots: Observe for data points that deviate significantly from the group pattern.

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.



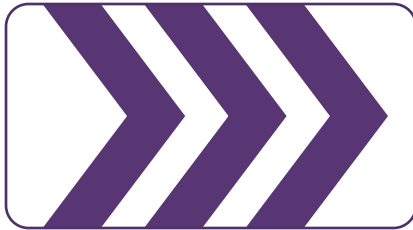
How to Correct Dirty Data?



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

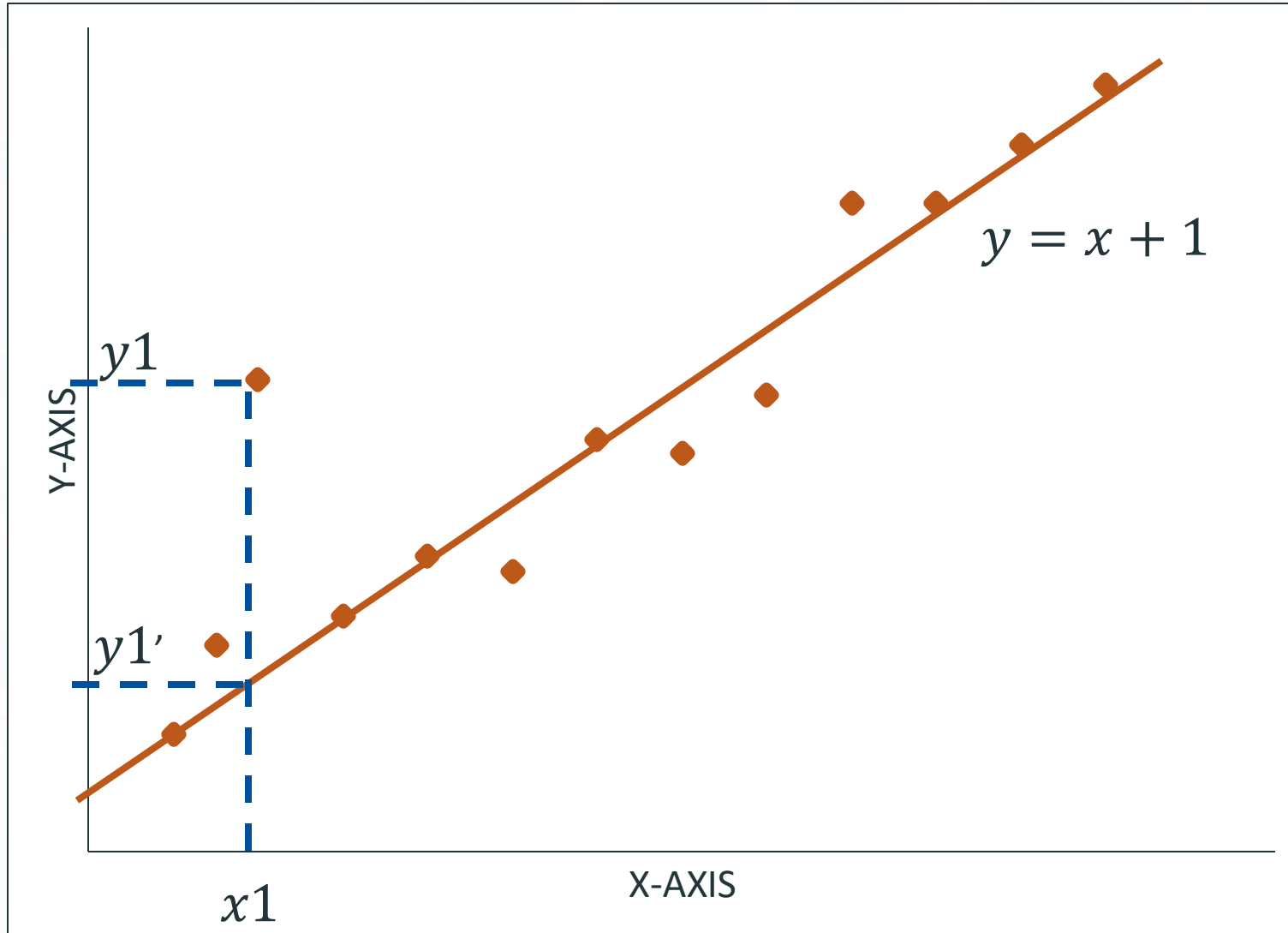
Smoothing by **bin means**:

- 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

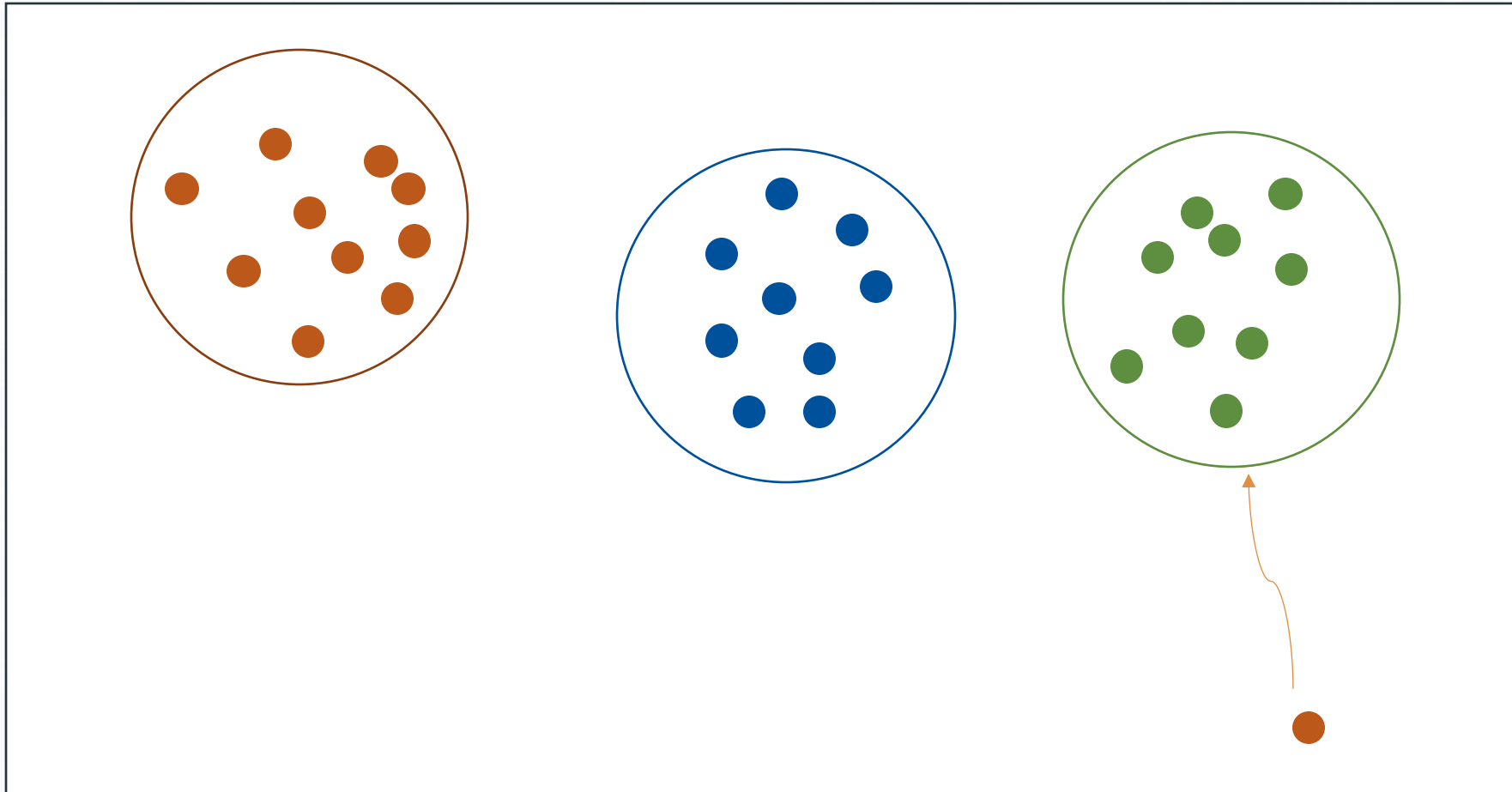
Regression For Data Smoothing



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 auxiliary 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.