Data Wrangling and Data Analysis
Data Preparation

Hakim Qahtan

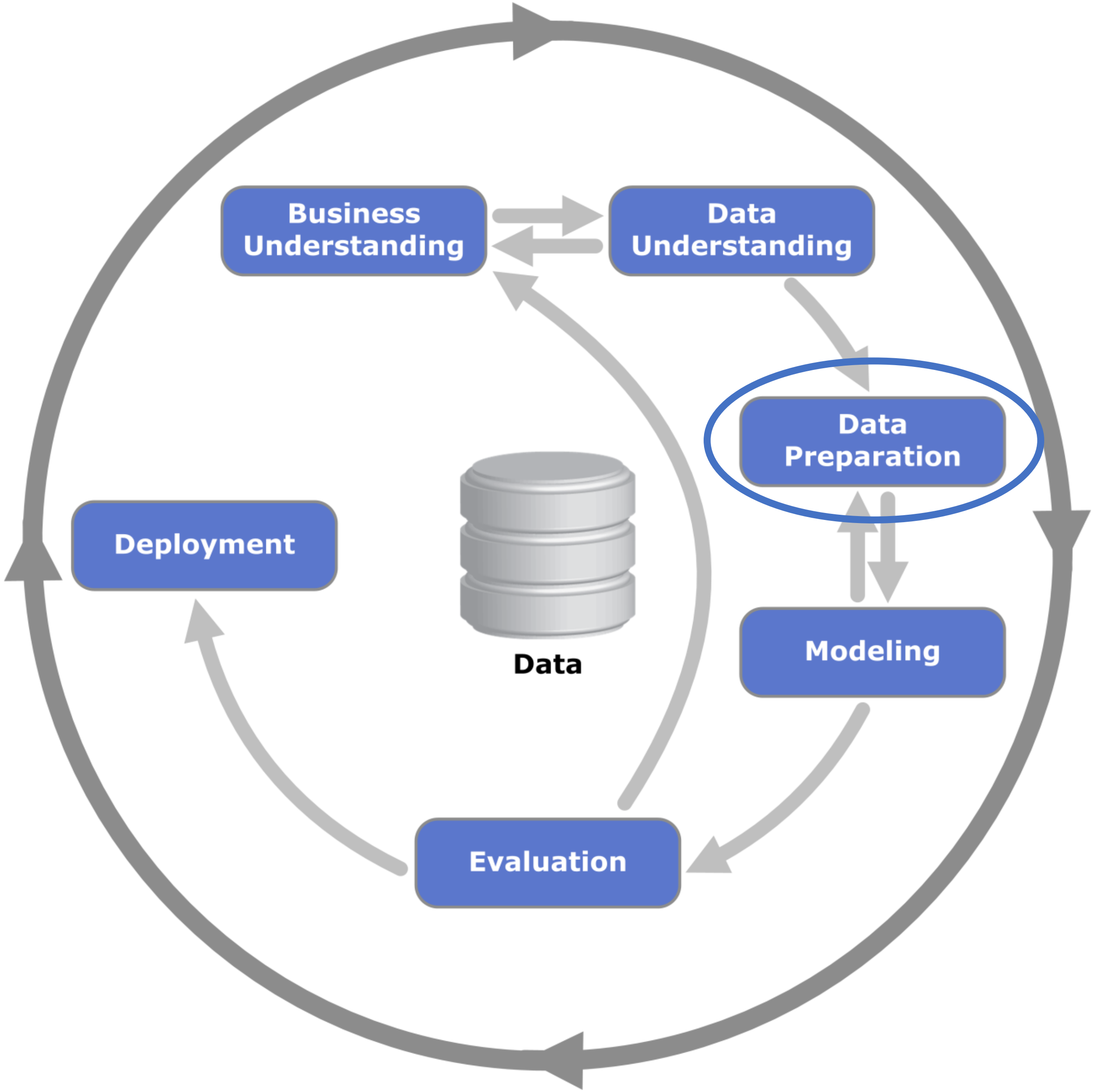
Department of Information and Computing Sciences

Utrecht University

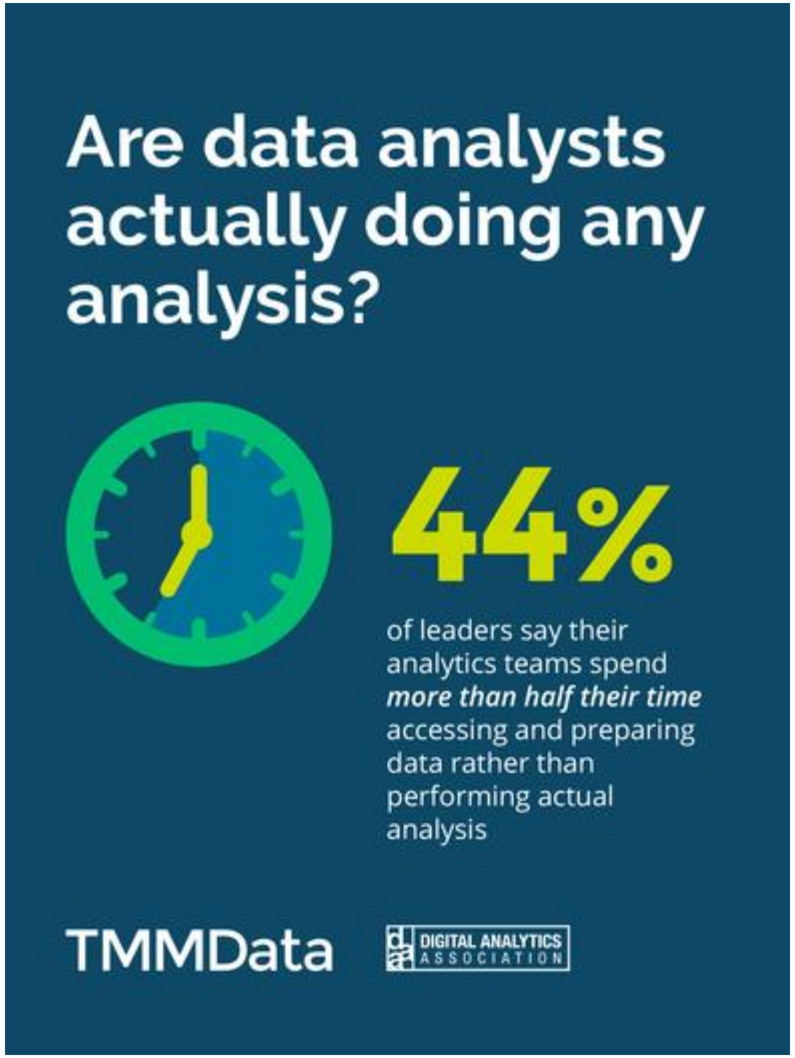
Topics for Today

- Data preparation: motivation and overview
- Quality of data
    - Data cleaning
        - Incomplete data
        - Noisy data
        - Outliers

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Deployment
Data
Modeling
Evaluation

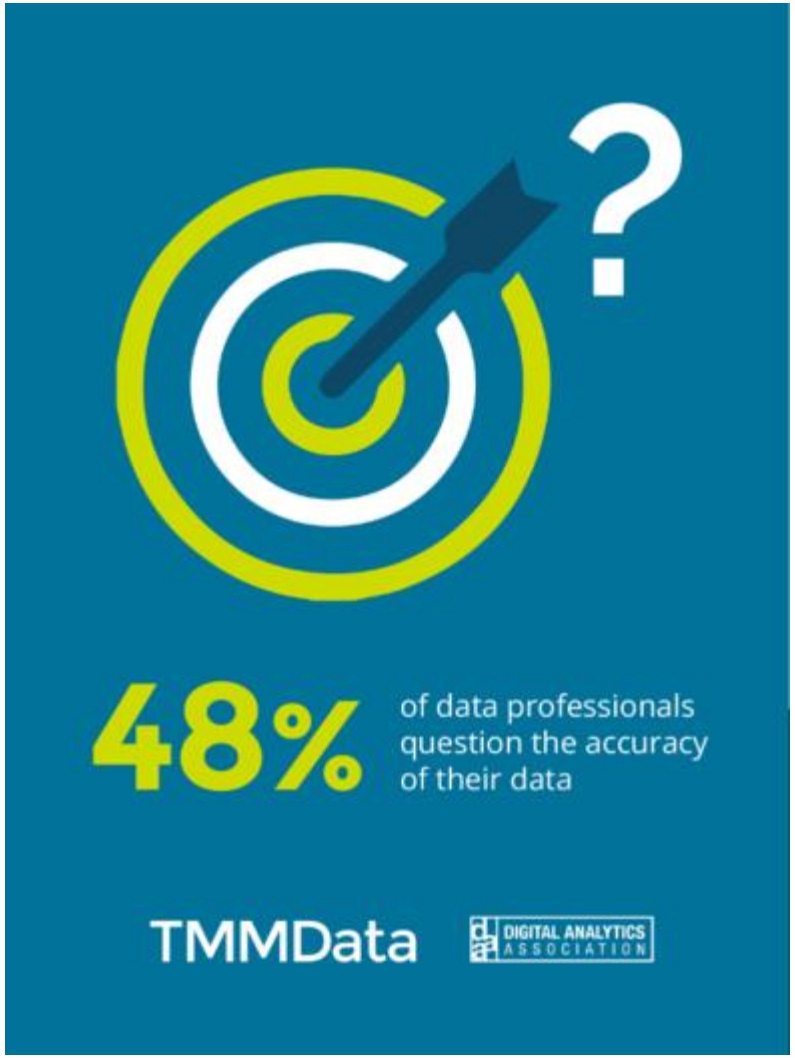
Why data preparation?



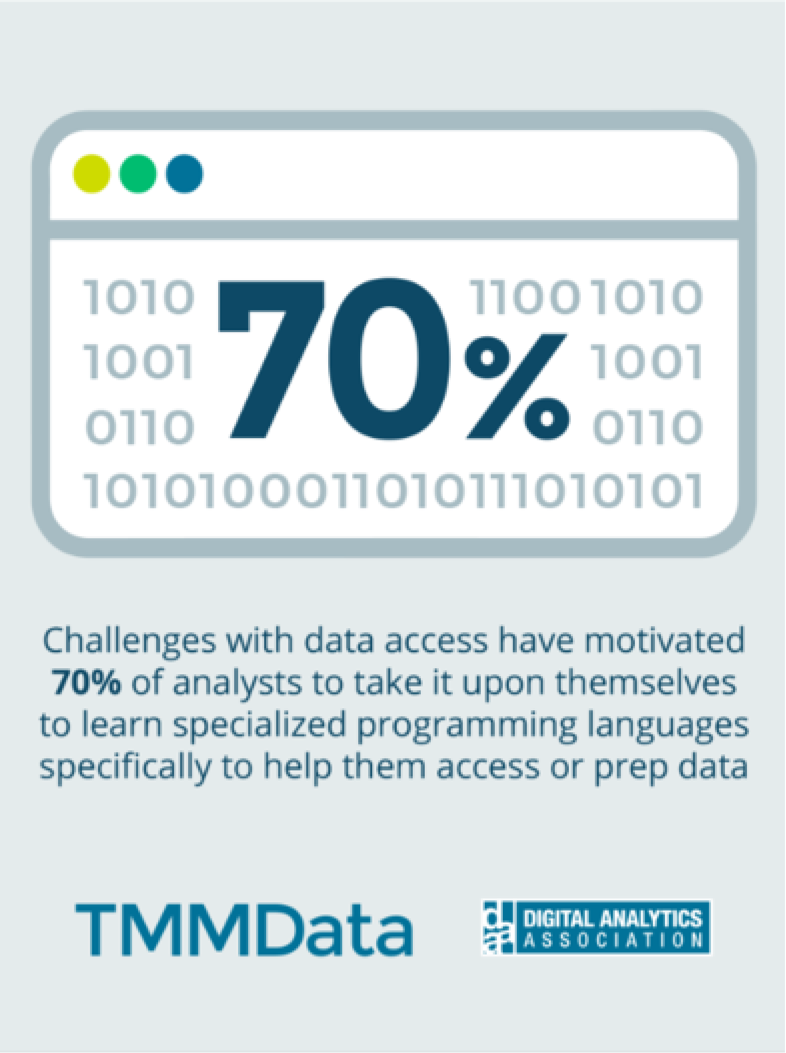
Are data analysts actually doing any analysis?

44%

of leaders say their analytics teams spend *more than half their time* accessing and preparing data rather than performing actual analysis



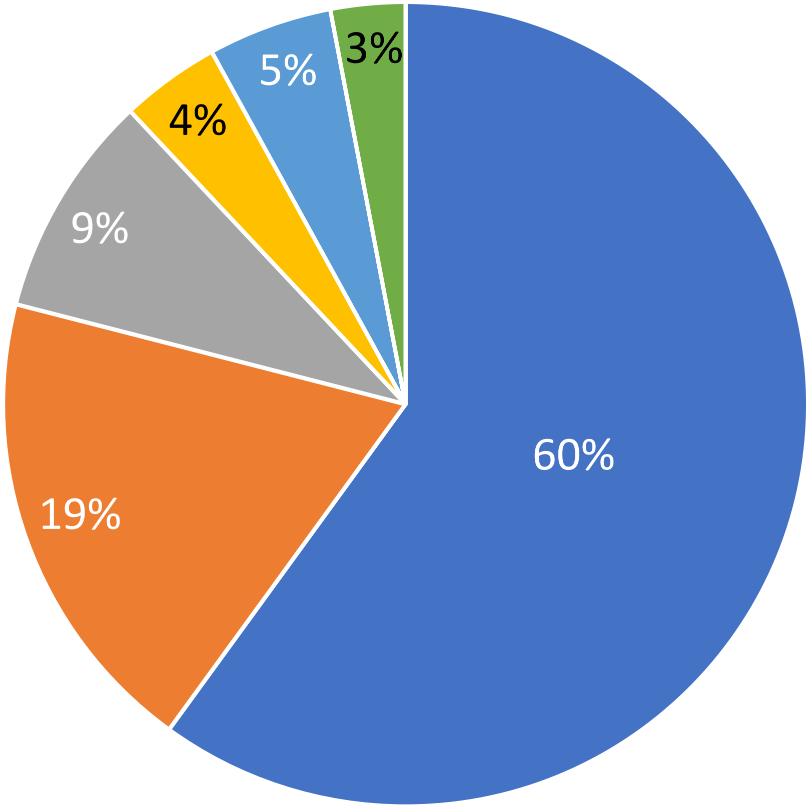
48% of data professionals question the accuracy of their data



Challenges with data access have motivated **70%** of analysts to take it upon themselves to learn specialized programming languages specifically to help them access or prep data

Why data preparation?

- What data scientists spend the most of their time doing:



🟦	Collecting datasets	60%
🟧	Cleaning and organizing the data	19%
⬜	mining data for patterns	9%
🟨	Refining algorithms	4%
🟦	Other	5%
🟩	Building training sets	3%

Data preparation: An Overview

- Data preparation: the process of cleaning and transforming raw data prior to processing and analysis.
- Involves the following tasks:
  - Checking the data quality and applying correction techniques
  - Reformatting data
  - Combining datasets for the purpose of enrichment

Data Quality

- Data quality: checking the data from multiple perspectives
    - Accuracy
        - The data was recorded correctly.
    - Completeness
        - All relevant data was recorded.
    - Uniqueness
        - Entities are recorded once.
    - Timeliness
        - The data is kept up to date.
            - Special problems in federated data: time consistency.
    - Consistency
        - The data agrees with itself.

Example

```
T.Das|97336o8237|24.95|Y|-|0.0|1000
Ted J.|973-360-8997|2000|N|M|NY|1000
```

- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field three is Revenue. In dollars or cents?

Data Quality (Cont.)

- Problems in previous definition of data quality:
    - Unmeasurable
        - Accuracy and completeness are extremely difficult, perhaps impossible to measure.
    - Context independent
        - No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.
    - Incomplete
        - What about interpretability, accessibility, metadata, analysis, etc.
    - Vague
        - The previous definition provide no guidance towards practical improvements of the data.

Data Quality (Cont.)

- Sources of problems:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (duplicates)
  - Approximations, surrogates – SW/HW constraints
  - Measurement errors.
- We need a definition of data quality which
  - Reflects the use of the data
  - Leads to improvements in processes
  - Measurable (we can define metrics)

Examples of Data Quality metrics

- Conformance to schema
  - Evaluate constraints on a snapshot.
- Conformance to business rules
  - Evaluate constraints on changes in the database.
- Accuracy
  - Perform inventory (expensive), or use proxy (track complaints). Audit samples?
- Glitches in analysis
- Successful completion of end-to-end process

Data Cleaning

Data Cleaning

- 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)

Data Cleaning (Cont.)

- Data in the real world is dirty
    - 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

Incomplete Data

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data not entered due to misunderstanding
  - Certain data values may not be considered important at the time of entry
- Missing data may need to be inferred

Handling the Missing Data

- Ignore the tuple: usually done when class label is missing (for classification problem) – not effective when the % of missing values is large
- Fill in the missing value:
  - Manually: tedious + infeasible?
  - Automatically (data imputation) with
    - A global constant : e.g., “unknown”, a new class?!
    - The attribute mean
    - The attribute mean for all samples belonging to the same class: smarter
    - More sophisticated data imputation techniques will be discussed in more details in the upcoming weeks

Data Cleaning

Noisy Data

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention

Handling the Noisy Data

- Binning
  - First sort data and partition into (equal-frequency) bins
  - Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
    - Smoothing by bin means : In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
    - Smoothing by bin median : In this method each bin value is replaced by its bin median value.
    - Smoothing by bin boundary: In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Handling the Noisy Data (Cont.)

- Regression
  - Smooth by fitting the data into regression functions
- Clustering
  - Detect and remove outliers that do not belong to any of the clusters
- Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Binning methods for data smoothing

Sorted data for grades (out of 50):

4, 8, 9, 11, 15, 21, 21, 24, 24, 25,26, 28, 29, 34, 48

Partition into equal-frequency (**equidepth**) bins:

- Bin 1: 4, 8, 9, 11, 15
- Bin 2: 21, 21, 24, 24, 25
- Bin 3: 26, 28, 29, 34, 48

Smoothing by **bin means**:

- Bin 1: 9.4, 9.4, 9.4, 9.4, 9.4
- Bin 2: 23, 23, 23, 23, 23
- Bin 3: 33, 33, 33, 33, 33

Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15, 15
- Bin 2: 21, 21, 25, 25, 25
- Bin 3: 26, 26, 26, 26, 48

Data Cleaning

Outliers

Outliers

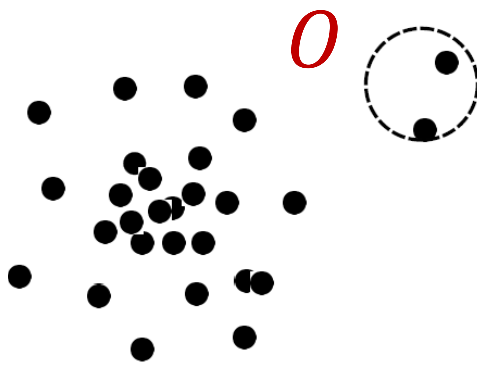
- Consider the data points

  3, 4, 7, 4, 8, 3, 9, 5, 7, 6, 92
- "92" is suspicious - an *outlier*
- Outliers are data or model glitches

Definition of Hawkins [Hawkins 1980]:
"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

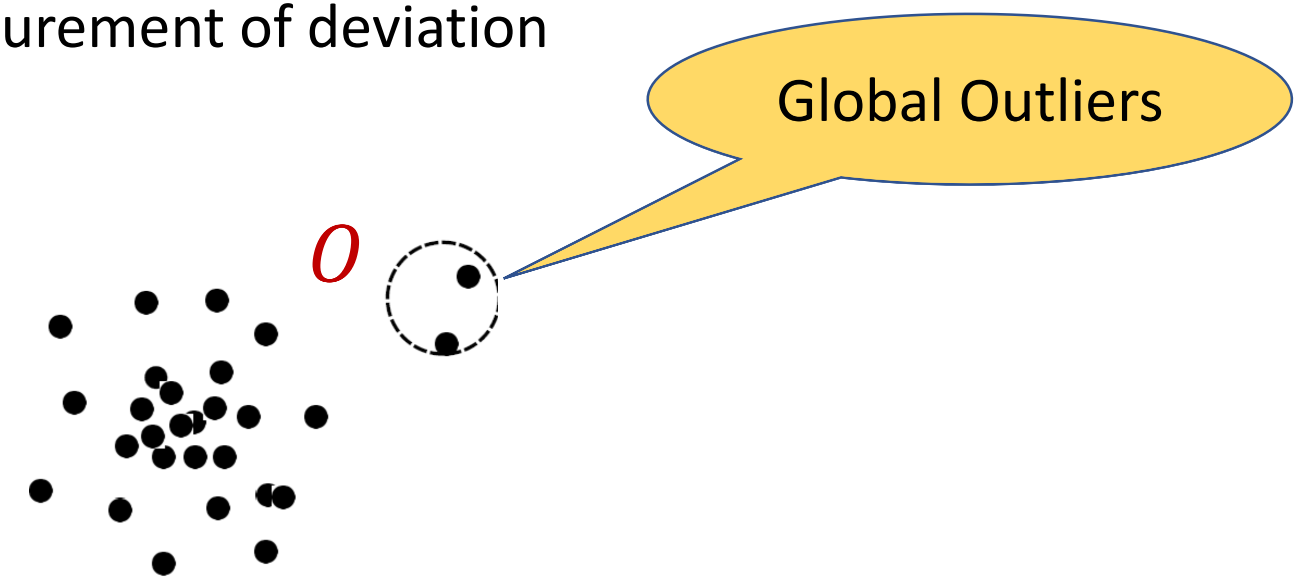
Outliers (Cont.)

- Outliers are different from the noise
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: they violate the mechanism that generates the normal data
- Outlier detection vs. *novelty detection*: early stage, outlier; but later merged into the model
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis



Types of Outliers

- Three kinds: *global, contextual* and *collective* outliers
- **Global outlier** (or point anomaly)
  - Object is a global outlier (o_g) if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation

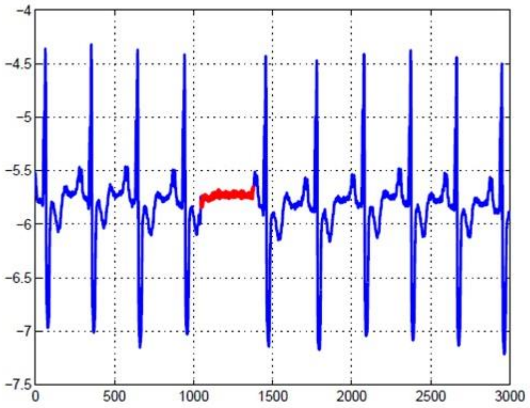


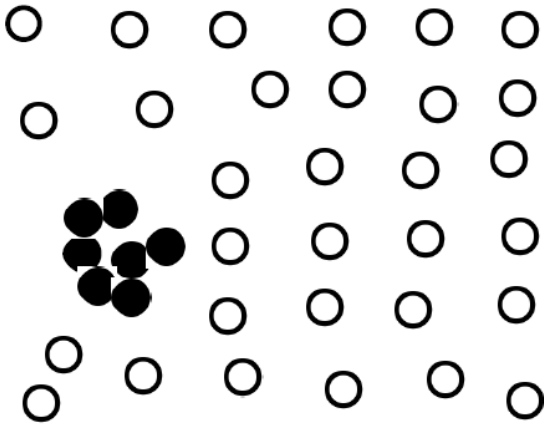
Types of Outliers (Cont.)

- **Contextual outlier** (or *conditional outlier*)
  - Object is o_c if it deviates significantly based on a selected context
  - Ex. 20^o C in Utrecht: outlier? (depending on summer or winter?)
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
  - Can be viewed as a generalization of *local outliers*—whose density significantly deviates from its local area
  - Issue: How to define or formulate meaningful context?

Types of Outliers (Cont.)

- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages to each other
  - Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects





Challenges of Outlier Detection

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection

Challenges of Outlier Detection (Cont.)

- Understandability
  - Understand why these are outliers: Justification of the detection
  - Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism
- A data set may have multiple types of outlier
- One object may belong to more than one type of outlier

Outlier Detection Techniques

- Many approaches, for example:
  - Statistical methods
  - Distance based methods
  - Density based methods
  - Methods based on model fitting
- Local vs global approaches
- Scoring vs labeling

Outlier Detection Techniques (Cont.)

- Global versus local approaches
  - Considers the resolution of the reference set w.r.t. which the “outlierness” of a particular data object is determined – Global approaches
- Global approaches
  - The reference set contains all other data objects
    - Basic problem: other outliers are also in the reference set and may falsify the results
- Local approaches
  - The reference contains a (small) subset of data objects
  - No assumption on the number of normal mechanisms
    - Basic problem: how to choose a proper reference set
- Some approaches are somewhat in between

Outlier Detection Techniques (Cont.)

- Labeling versus scoring: considers the output of an outlier detection algorithm
  - Labeling approaches
    - Binary output
    - Data objects are labeled either as normal or outlier
  - Scoring approaches
    - Continuous output
    - For each object an outlier score is computed (e.g. the probability for being an outlier)
    - Data objects can be sorted according to their scores

Outlier Detection – Statistical Approaches

- Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model)
- Idea: learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers
- Methods are divided into two categories: *parametric* vs. *non-parametric*

Outlier Detection – Statistical Approaches (Cont.)

- Parametric method
  - Assumes that the normal data is generated by a parametric distribution with parameter θ
  - The probability density function of the parametric distribution $f(x, \vartheta)$ gives the probability that object x is generated by the distribution
  - The smaller this value, the more likely x is an outlier
- Non-parametric method
  - Not assume an a-priori statistical model and determine the model from the input data
  - Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance
  - Examples: histogram and kernel density estimation

Outlier Detection – Statistical Approaches (Cont.)

- Detection of univariate outliers based on Normal distribution
- Univariate data: A data set involving only one attribute or variable
- Often assume that data are generated from a normal distribution, learn the parameters from the input data, and identify the points with low probability as outliers

Outlier Detection – Statistical Approaches (Cont.)

- Ex: Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4, 29.5}
  - Use the maximum likelihood method to estimate μ and σ

$$\ln\left(\mathcal{L}(\mu, \sigma^2)\right) = \sum_{i=1}^{n} \ln\left(f(x_i|(\mu, \sigma^2))\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - Taking derivatives with respect to μ and σ^2, we derive the following maximum likelihood estimates

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - For the above data with n=10, we have $\hat{\mu} = 28.61$ and $\hat{\sigma} = \sqrt{2.24}$ = 1.497 Then (24 – 28.61) /1.497 = – 3.15 < –3, 24 is an outlier since for Normal distribution, $\hat{\mu} \pm 3\sigma$ region contains 99.7% of the data.

Outlier Detection – Statistical Approaches (Cont.)

- Grubb’s test (or maximum normed residual test)
  - Detecting outlier for univariate data under normal distribution assumption
  - For each object x in the dataset, compute the z-score

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}}$$

where $t^2_{\frac{\alpha}{2N},N-2}$ is the value taken by a t-distribution at a significance level of $\frac{\alpha}{2N}$, and N is the number of objects in the dataset

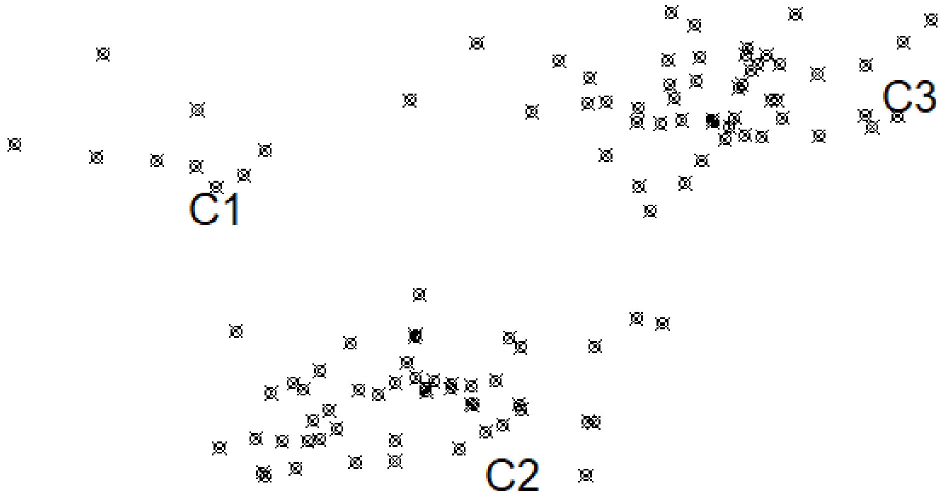
- Univariate data – data with single dimension (feature)
- Review the t-statistics (t-distribution)

Outlier Detection – Statistical Approaches (Cont.)

- **Multivariate data**: A data set involving two or more attributes or variables
- Transform the multivariate outlier detection task into a univariate outlier detection problem
- **Method 1**. Compute Mahalanobis distance
  - Let $\bar{o}$ be the mean vector for a multivariate data set. Mahalanobis distance for an object o to $\bar{o}$ is $\mathrm{MDist}(o, \bar{o}) = (o - \bar{o})^T S^{-1}(o - \bar{o})$ where S is the covariance matrix
  - Use the Grubb's test on this measure to detect outliers

Outlier Detection – Statistical Approaches (Cont.)

- Assuming data generated by a normal distribution could be sometimes overly simplified
- Example (data in the figure): The objects between the two clusters cannot be captured as outliers since they are close to the estimated mean
- To overcome this problem, assume the normal data is generated by mixture of Normal distributions.



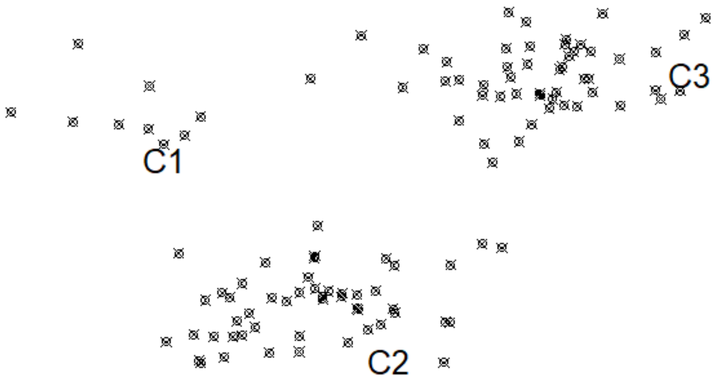
Outlier Detection – Statistical Approaches (Cont.)

- For any object o in the data set, the probability that o is generated by the mixture of the two distributions is given by

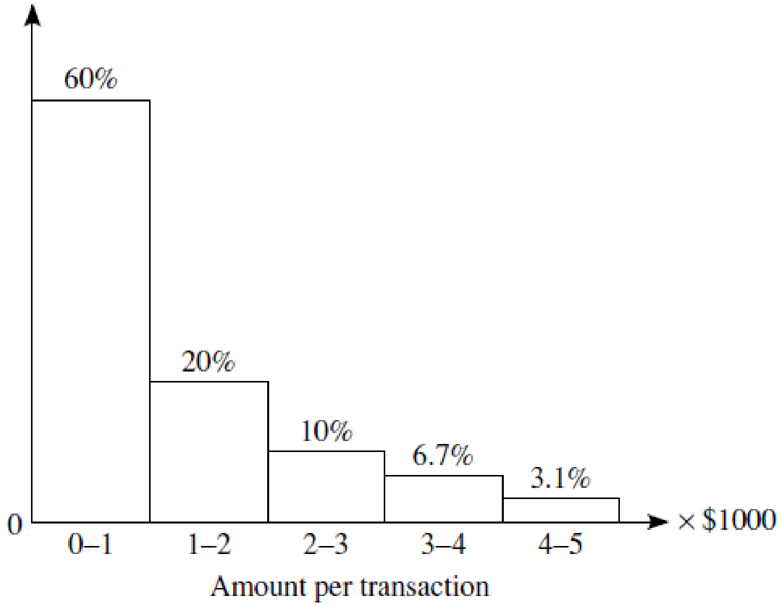
$$\Pr(o|\Theta_1, \Theta_2, \Theta_3) = f_{\Theta_1}(o) + f_{\Theta_2}(o) + f_{\Theta_3}(o)$$

where $f_{\Theta_1}(o), f_{\Theta_2}(o), f_{\Theta_3}(o)$are the probability density functions of $\Theta_1, \Theta_2, \Theta_3$

- use EM algorithm to learn the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3$ from data
- An object o is an outlier if it does not belong to any of the main groups of the data

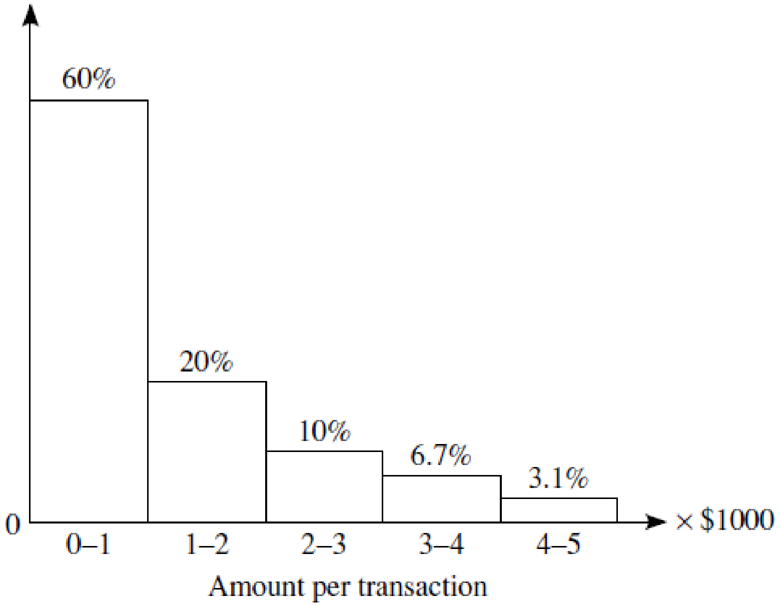


Outlier Detection – Statistical Approaches (Cont.)



- The model of normal data is learned from the input data without any *a priori* structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram:
  - Figure shows the histogram of purchase amounts in transactions
  - A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000
- Problem: Hard to choose an appropriate bin size for histogram

Outlier Detection – Statistical Approaches (Cont.)



- Problem: Hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, false positive
  - Too big bin size → outliers in some frequent bins, false negative
- Solution: Adopt kernel density estimation to estimate the probability density distribution of the data. If the estimated density function is high, the object is likely normal. Otherwise, it is likely an outlier.

Outlier Detection – Distance-Based Approaches

- General Idea
  - Judge a point based on the distance(s) to its neighbors
  - Several variants proposed
- Basic Assumption
  - Normal data objects have a dense neighborhood
  - Outliers are far apart from their neighbors, i.e., have a less dense neighborhood
- Example: DB $DB(\epsilon, \pi)$-Outliers
  - Idea: given a radius ϵ and a percentage π, a data point x is considered outlier if at most π of all other data points have distance less than ϵ to x
  - $outliers(\epsilon, \pi) = \{x \mid \frac{|\{y \in DB \mid dist(x,y) < \epsilon\}|}{|DB|} \leq \pi$

Outlier Detection – Density-Based Approaches

- General Idea
  - Compare the density around a point with the density around its local neighbors
  - The relative density of a point compared to its neighbors is computed as an outlier score
  - Approaches essentially differ in how to estimate density
- Basic Assumption
  - The density around a normal data object is similar to the density around its neighbors
  - The density around an outlier is considerably different to the density around its neighbors

Outlier Detection – Density-Based Approaches (Cont.)

- Local outlier factor (LOF):
  - Quantifies the local density of a data point, with the use of a neighborhood of size k
  - Introduces a smoothing parameter: Reachability-Distance RD

$$RD_k(x,y) = max\{Kdist(x), dist(x,y)\}$$

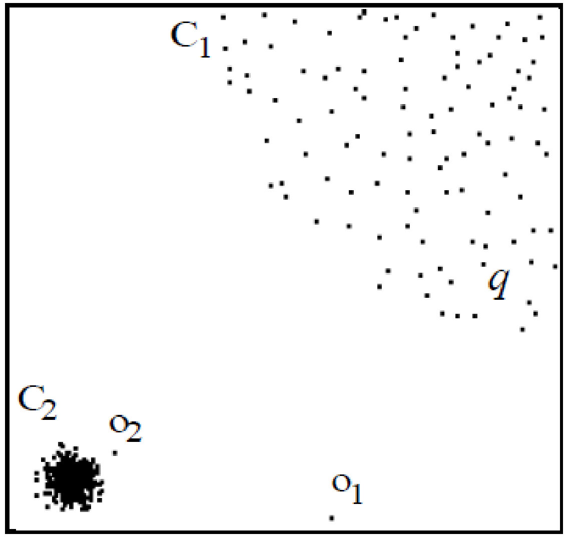
$Kdist(x)$ is the distance to the K-th neighbor of x

  - Local reachability distance (LRD) of a data point x is:

$$LRD_k(x) = 1/\left(\frac{\sum_{y\in kNN(x)} RD_k(x,y)}{k}\right)$$

  - The LOF is computed as

$$LOF_k(x) = \left(\frac{\sum_{y\in kNN(x)} \frac{LRD_k(y)}{LRD_k(x)}}{k}\right)$$



Outlier Detection – Model-Based Approaches

- Models summarize general trends in data
  - More complex than simple aggregates
  - E.g. linear regression, logistic regression
- Data points that do not conform to the fitting model are *potential outliers*
- Goodness of fit tests (DQ for analysis/mining)
  - Check suitableness of model to data
  - Verify validity of assumptions
  - Data rich enough to answer analysis/business question?

Reading Material & Exercises

- Chapters 3, 12 of the Data Mining: Concepts and Techniques Book