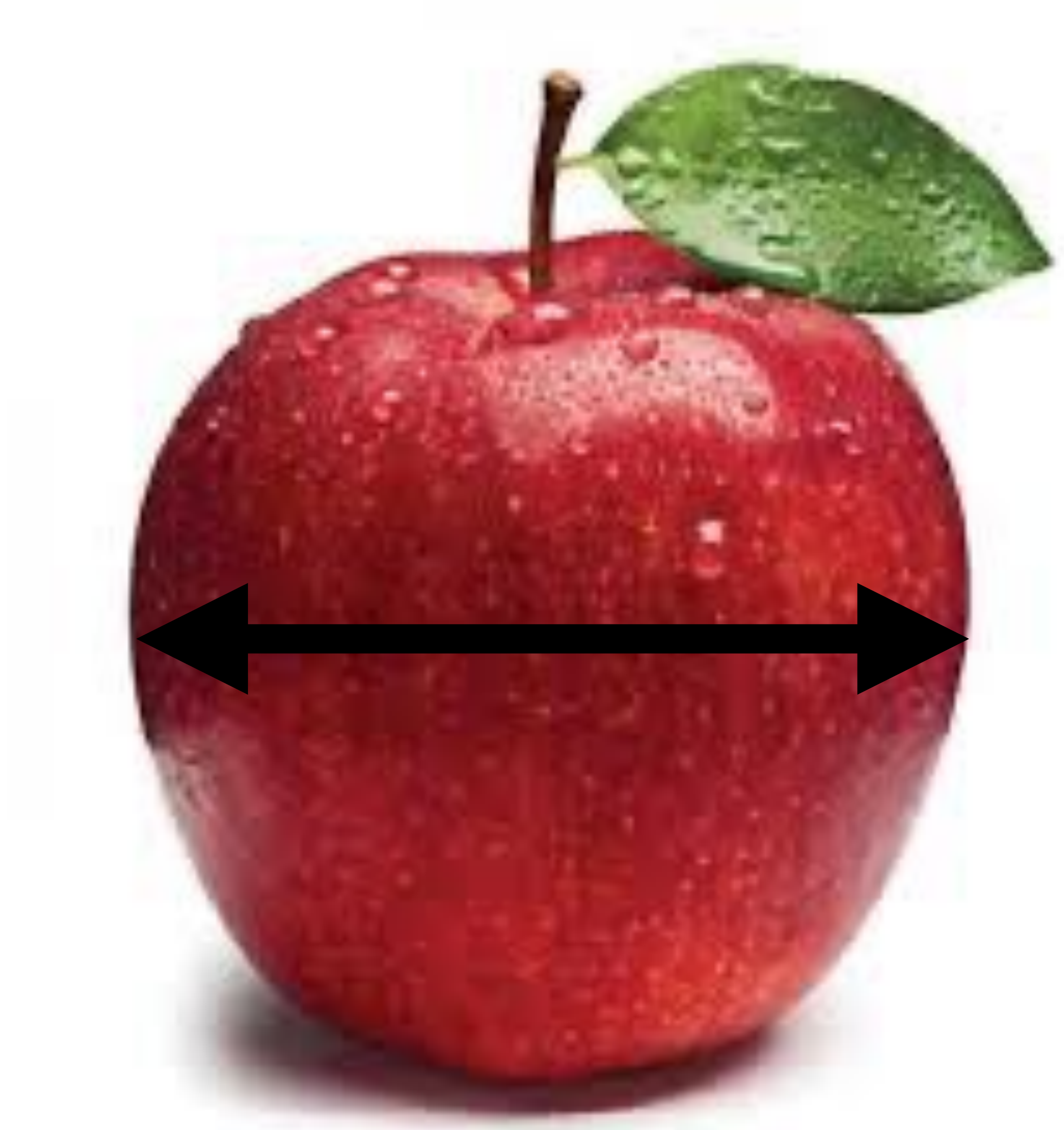
Introduction to Data Science and Programming, Fall 2019

Class 12: Single variable data analysis

Instructor: Michael Szell
Oct 4, 2019

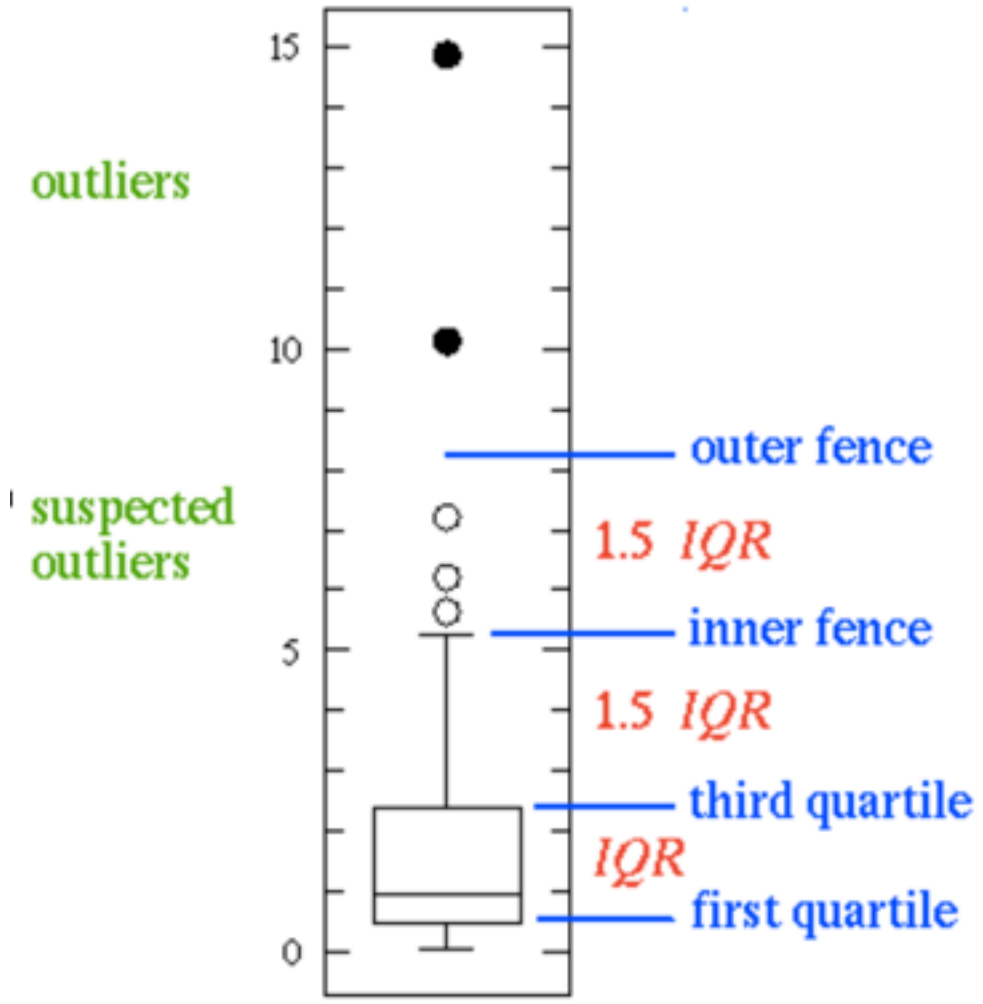


IT UNIVERSITY OF COPENHAGEN

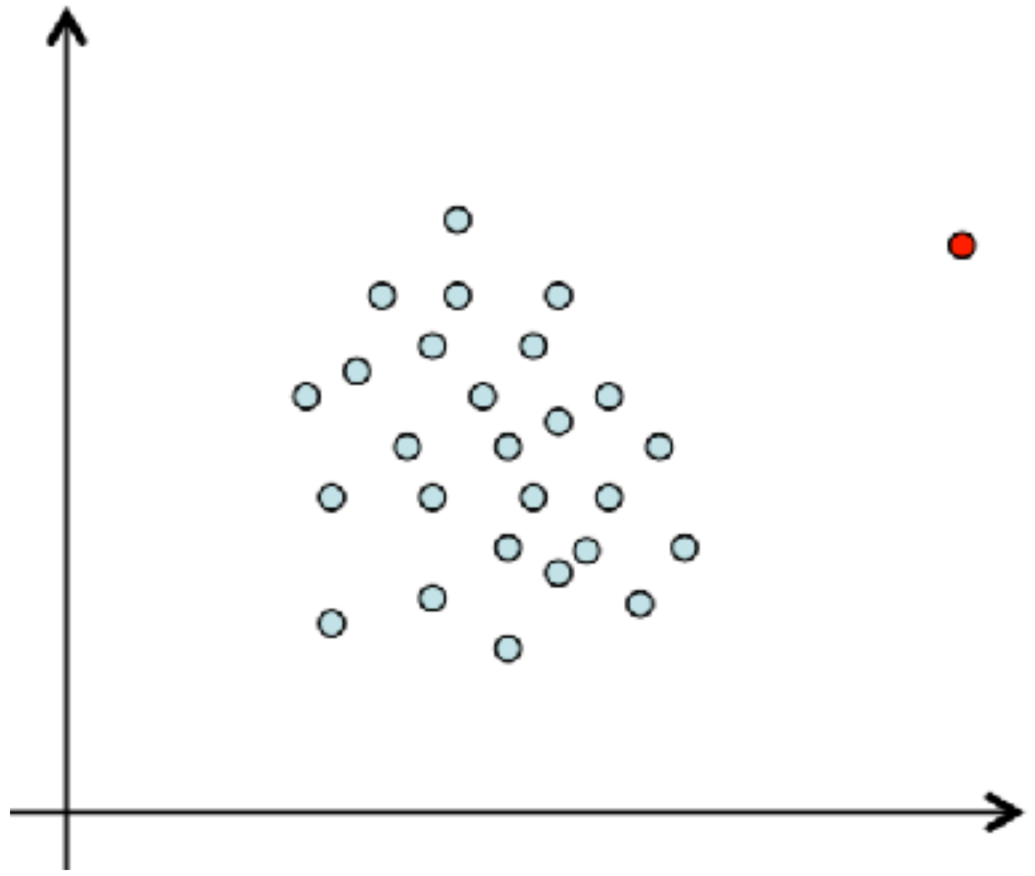
Today you will learn first steps in analyzing data

Variable types

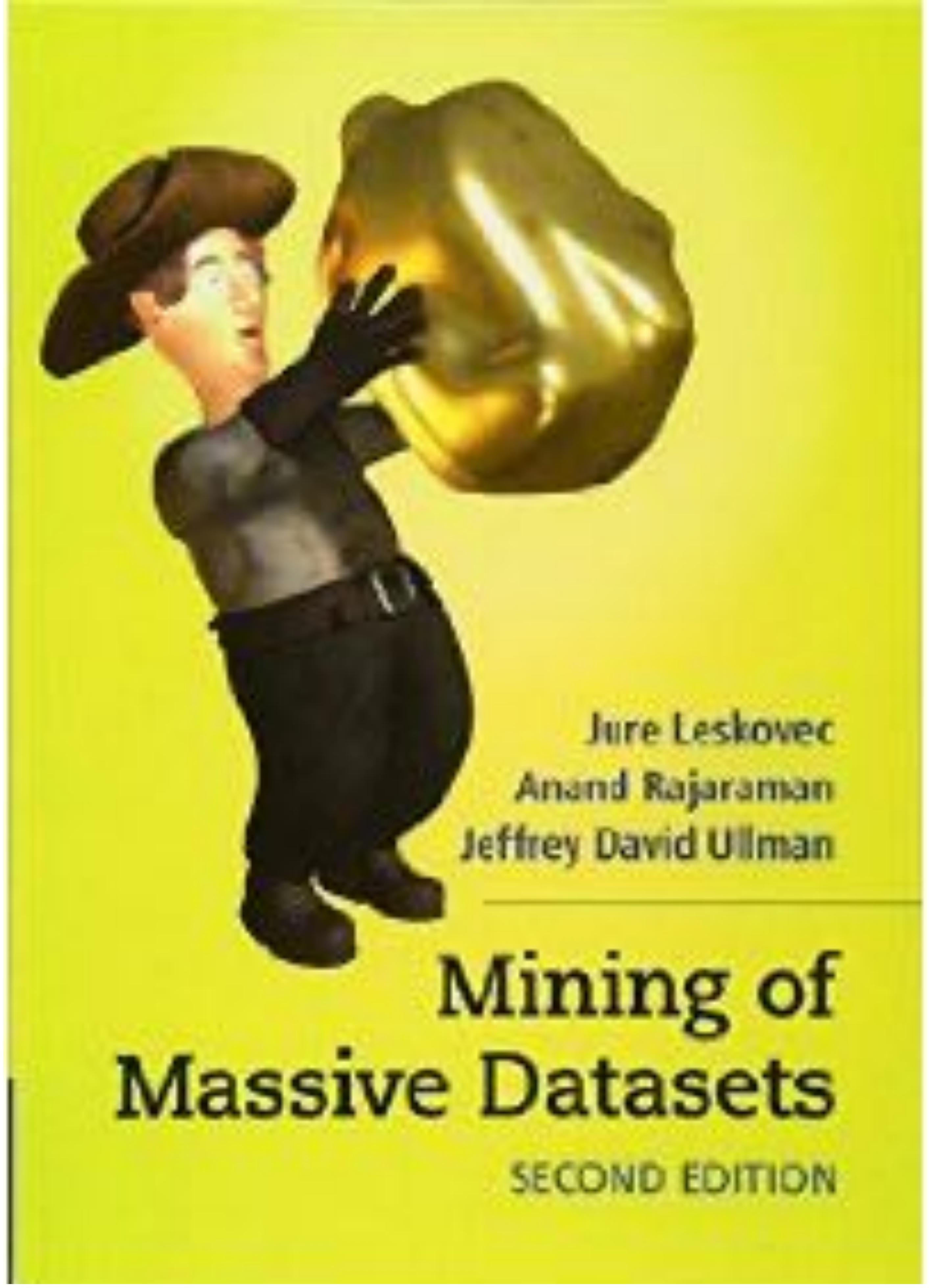
Describing data



Exploratory data analysis



We need data analysis to create knowledge from data



Jure Leskovec
Anand Rajaraman
Jeffrey David Ullman

Mining of Massive Datasets

SECOND EDITION

Computerization produces massive amounts of data

Knowledge discovered from data can be used for

- competitive advantage
- scientific advances

We need smart, automatized tools to deal with the massive data

We are drowning in data but starving for knowledge

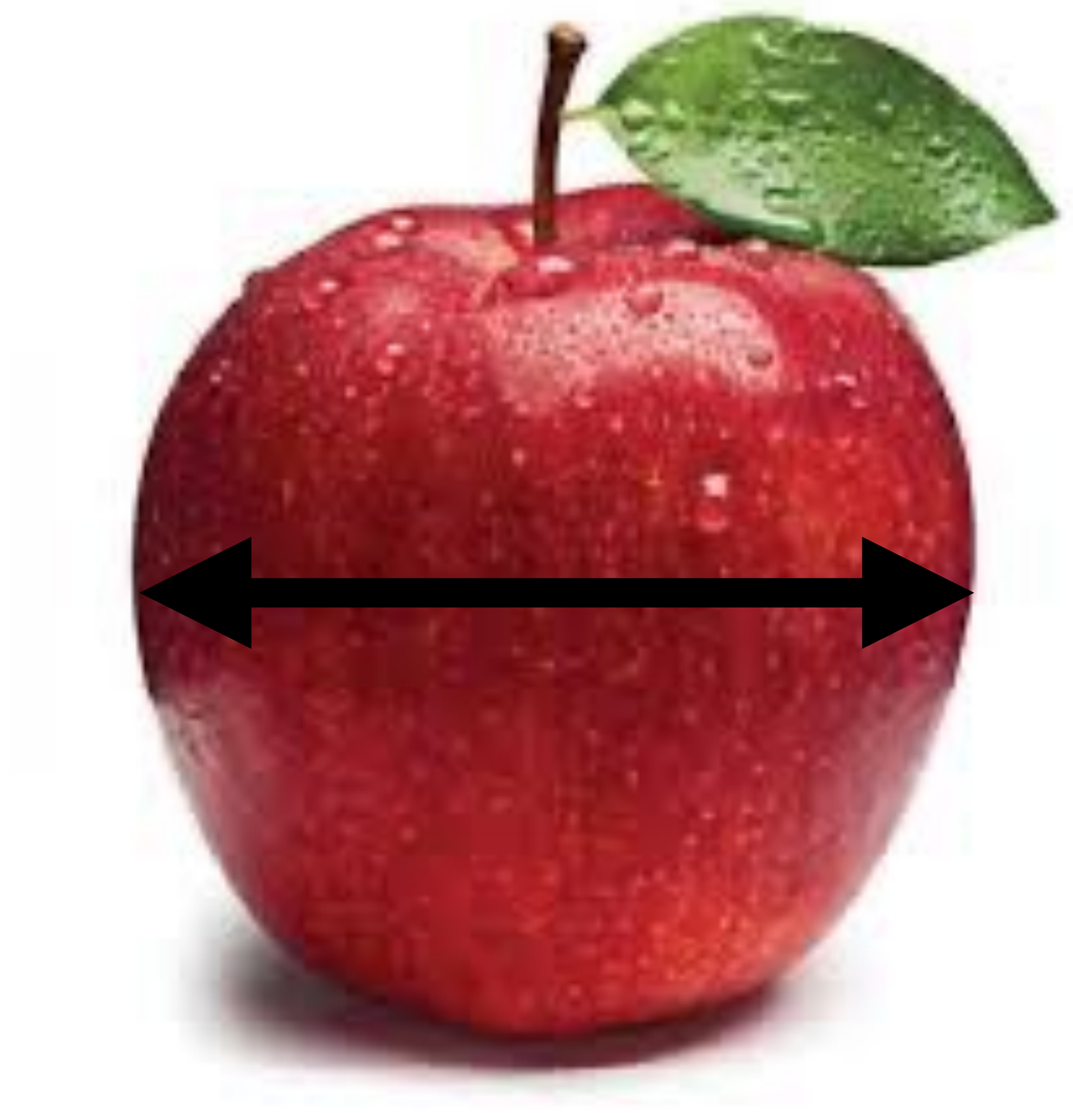
Data analysis is the process of:

Cleaning, transforming, exploring and/or modeling data

with the goal of discovering useful information,
informing conclusions or decision-making

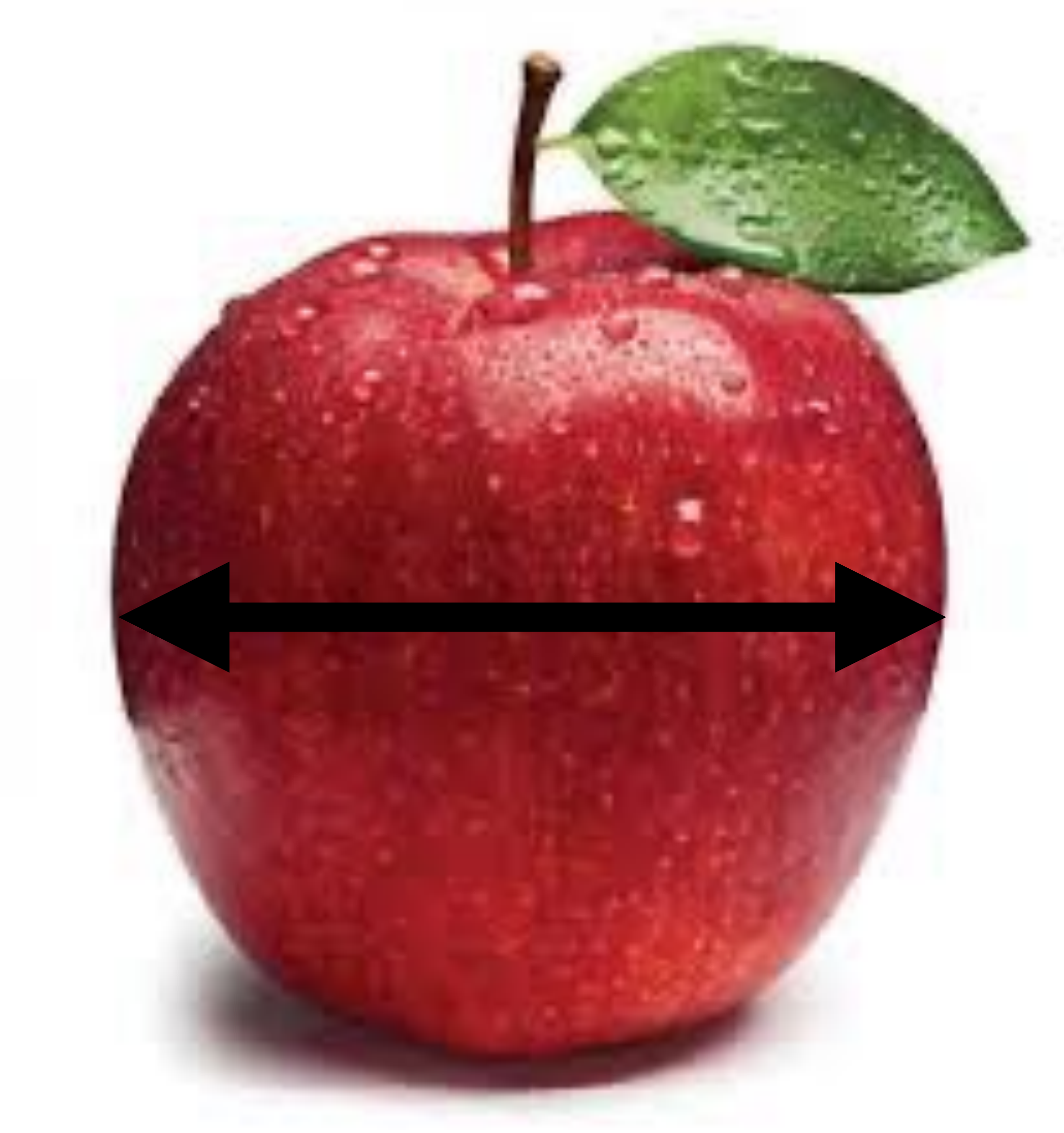
There are 3 types of data analysis

1) Descriptive statistics

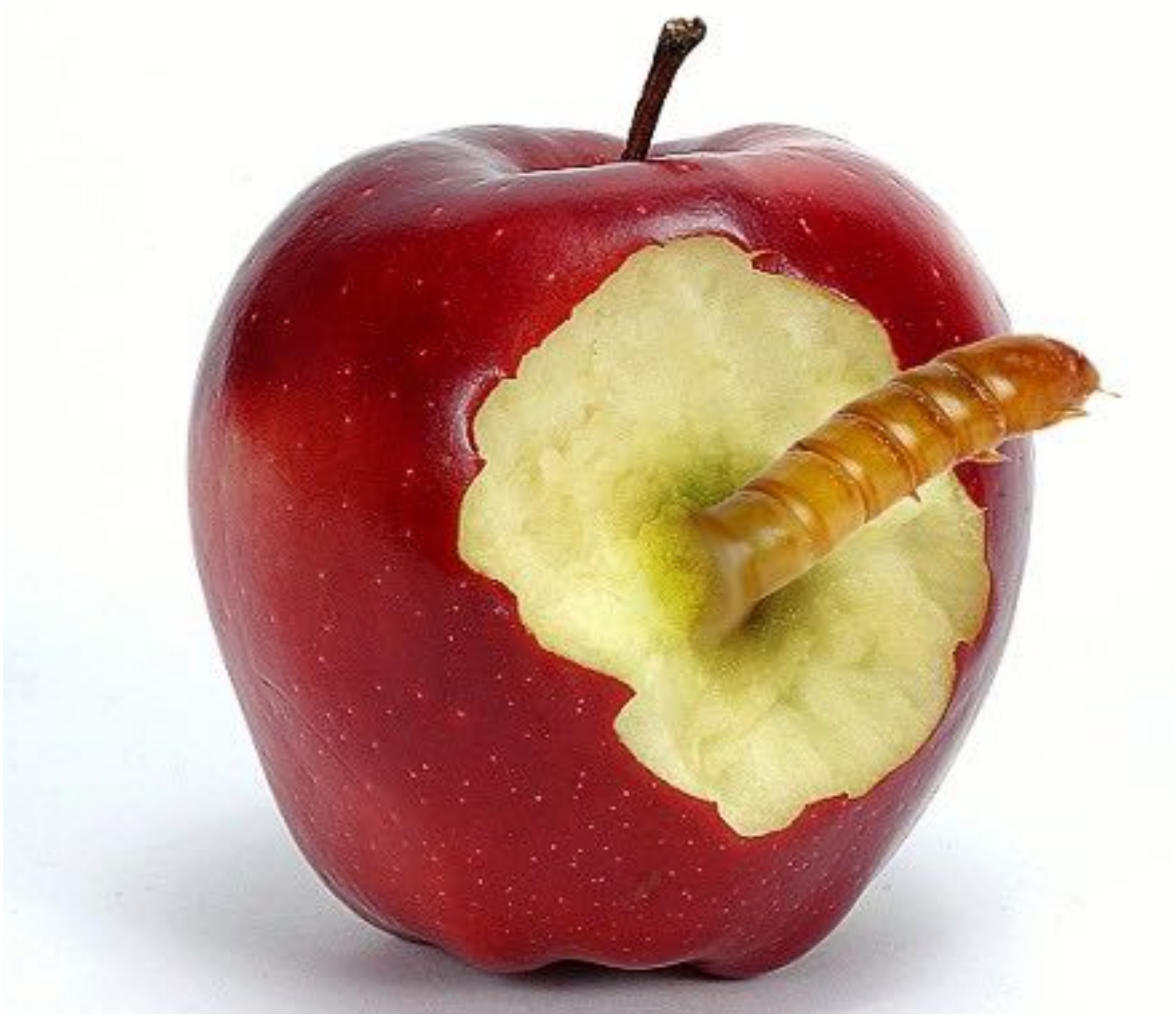


There are 3 types of data analysis

1) Descriptive statistics

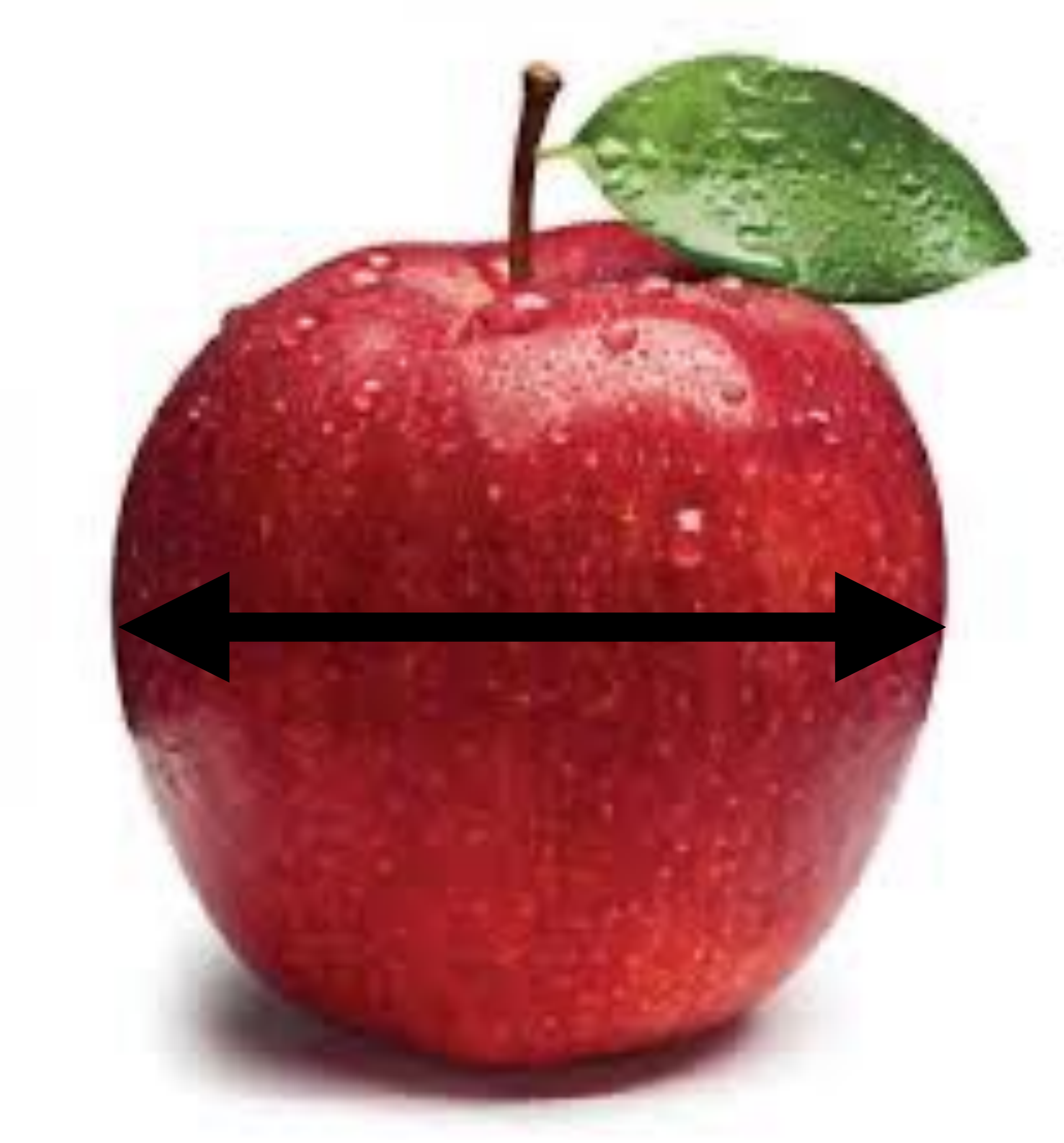


2) Exploratory

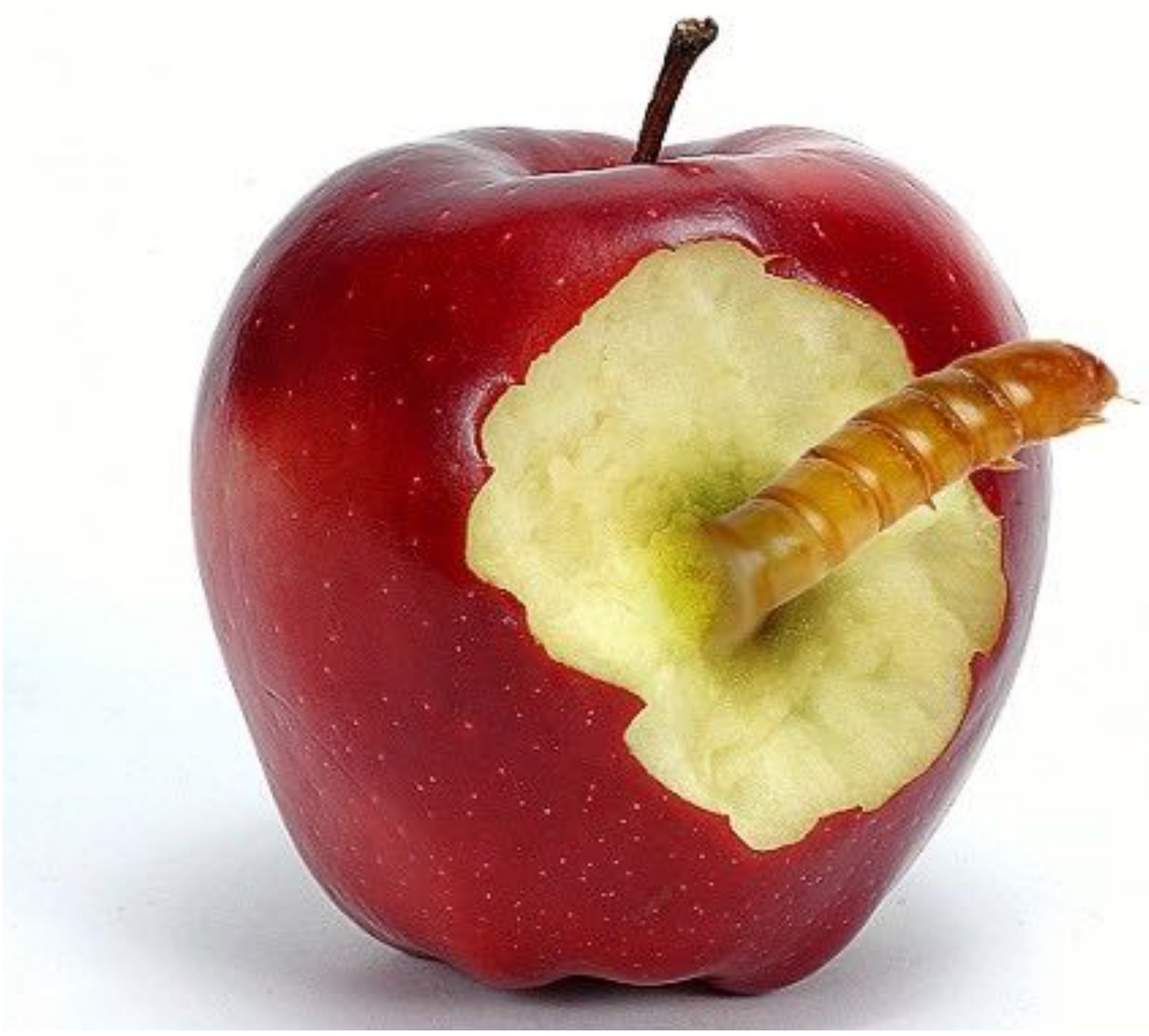


There are 3 types of data analysis

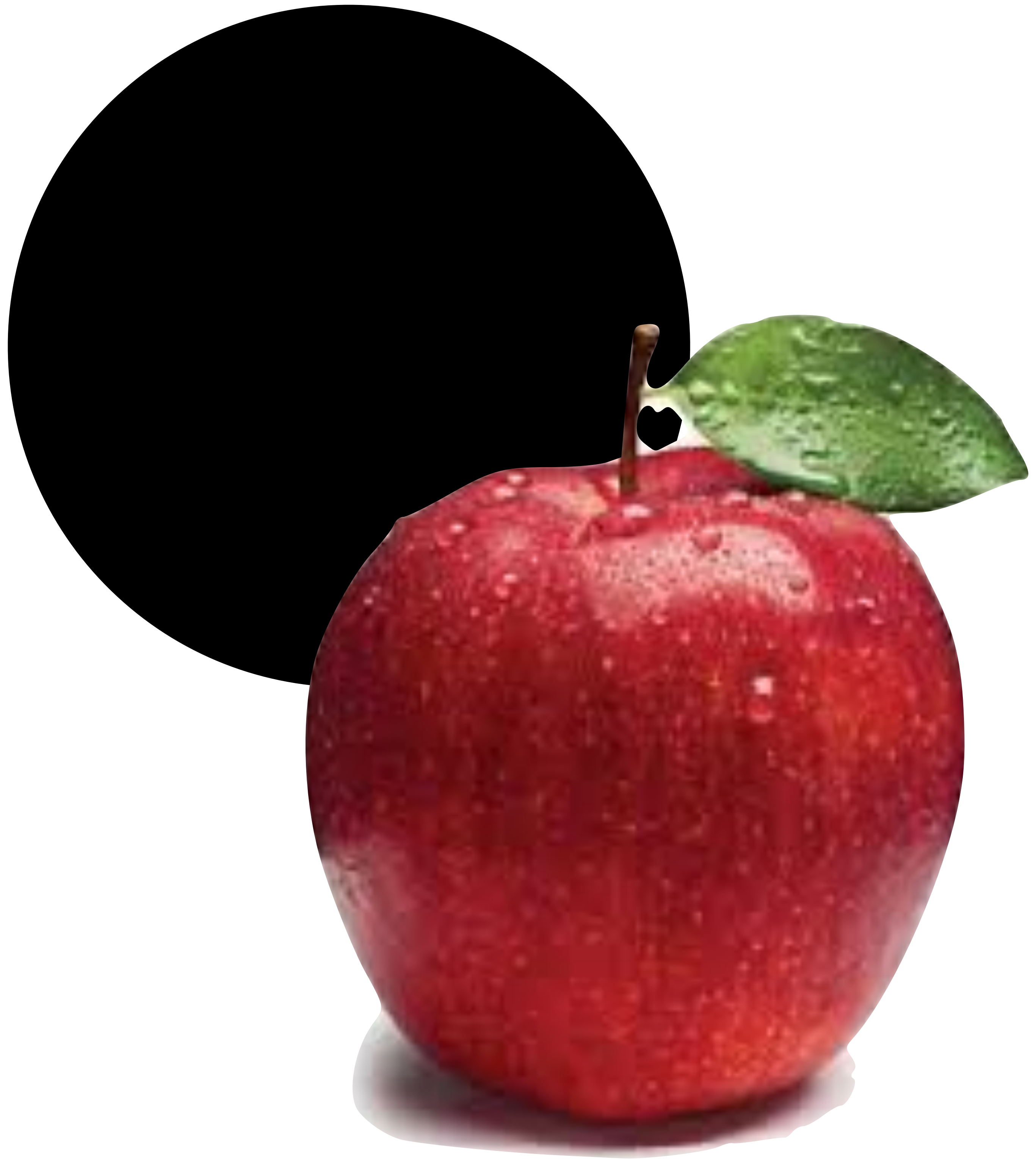
1) Descriptive statistics



2) Exploratory

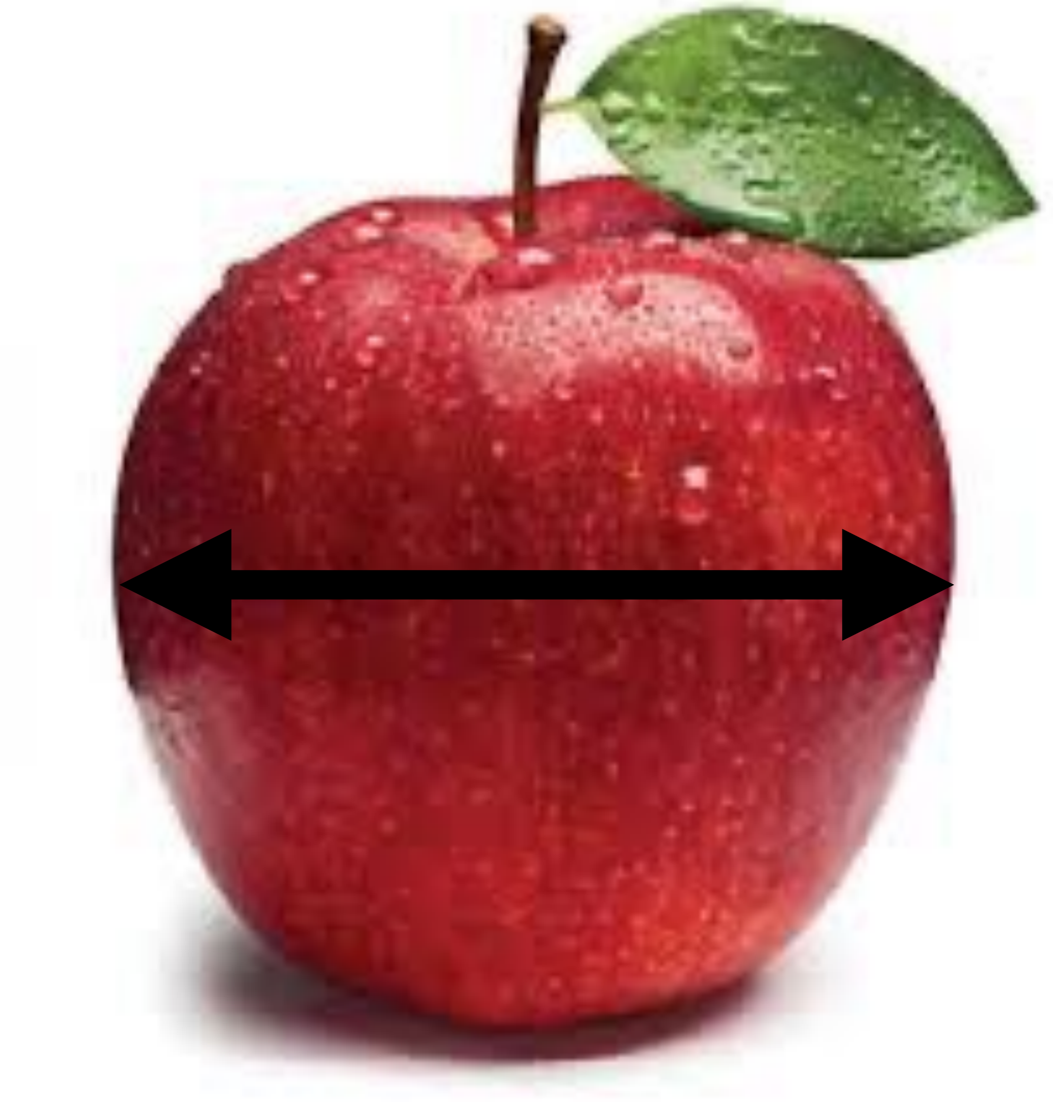


3) Inferential statistics
(Hypothesis testing)

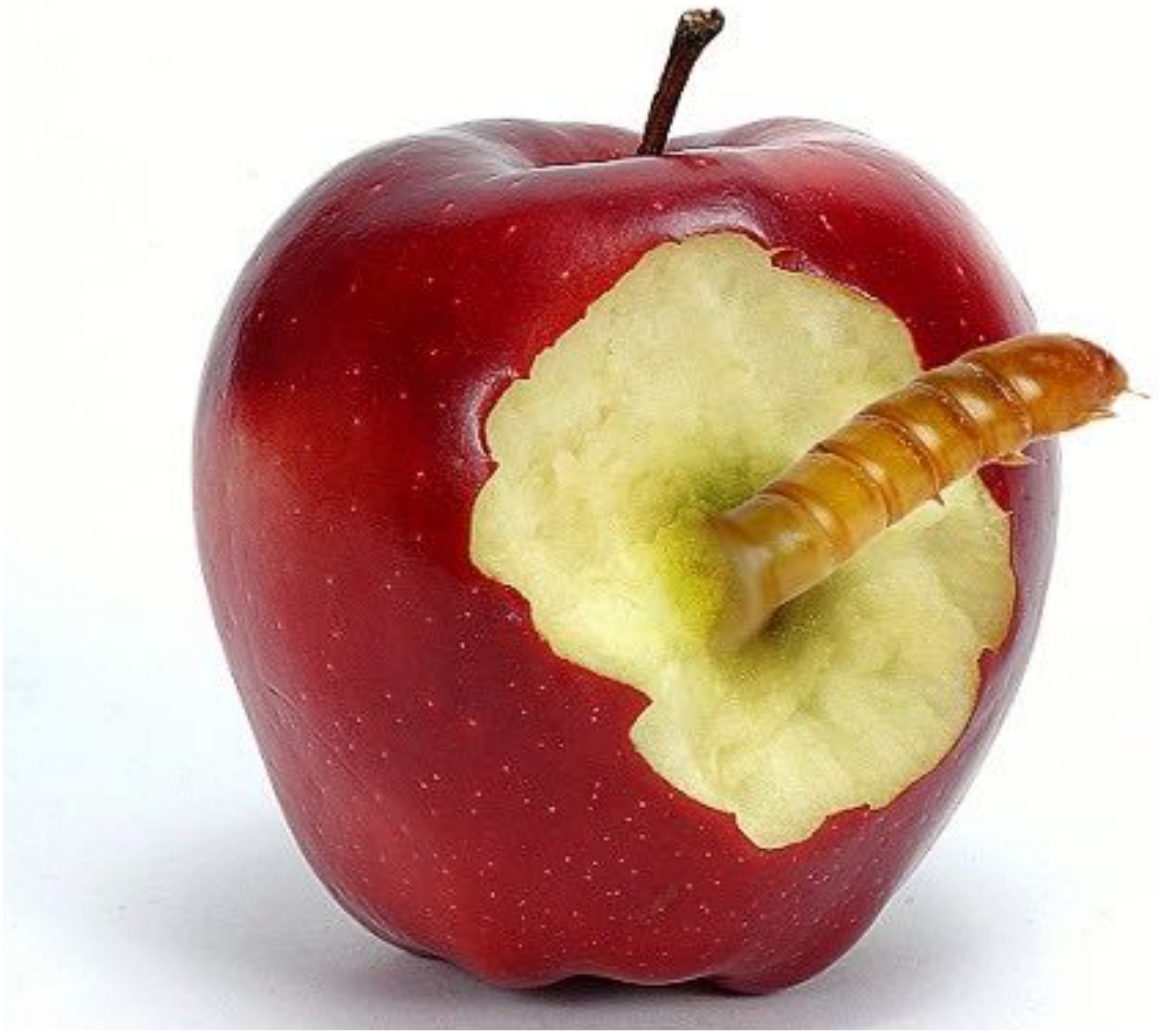


Today we focus on descriptive statistics and exploration

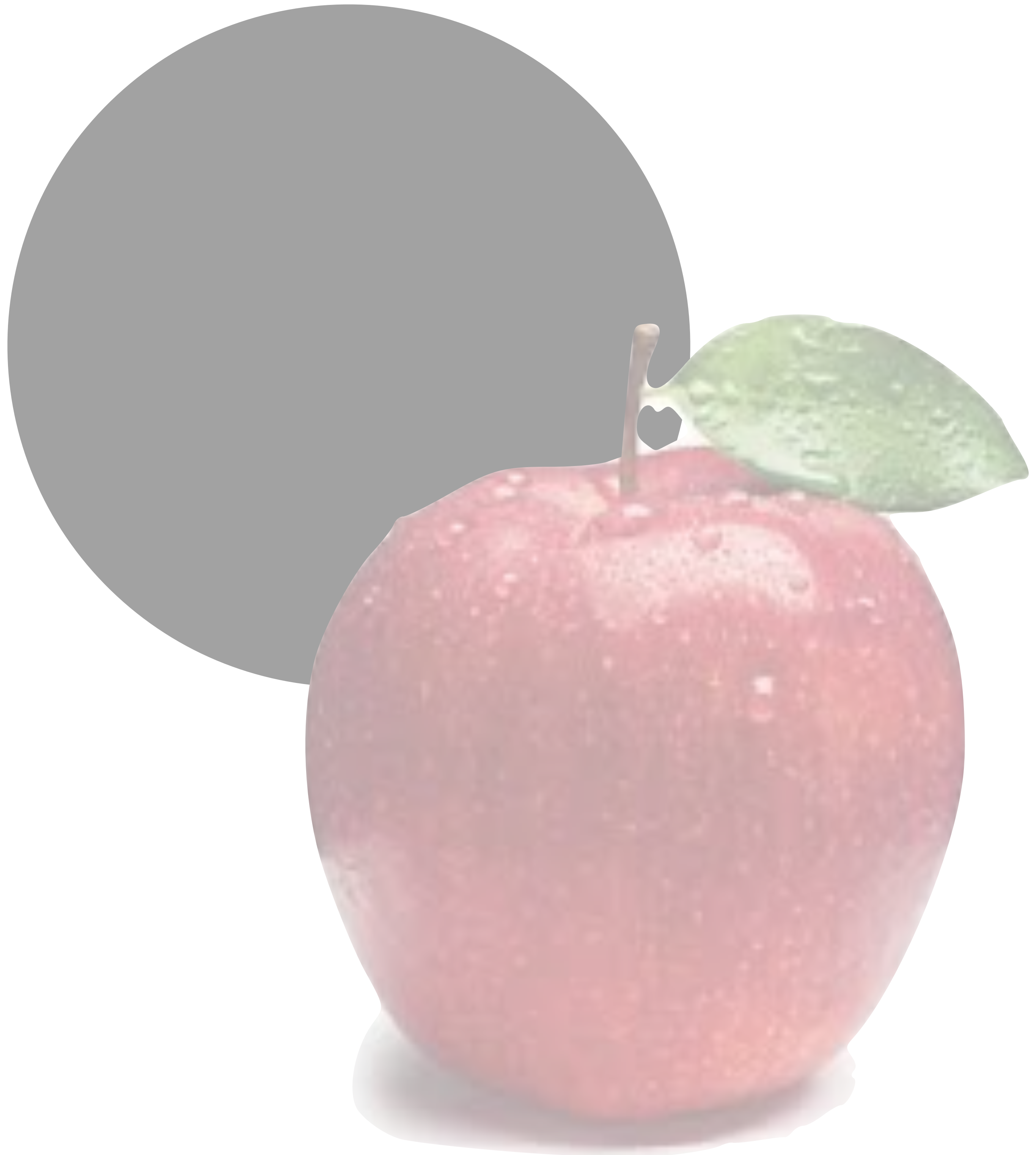
1) Descriptive statistics



2) Exploratory



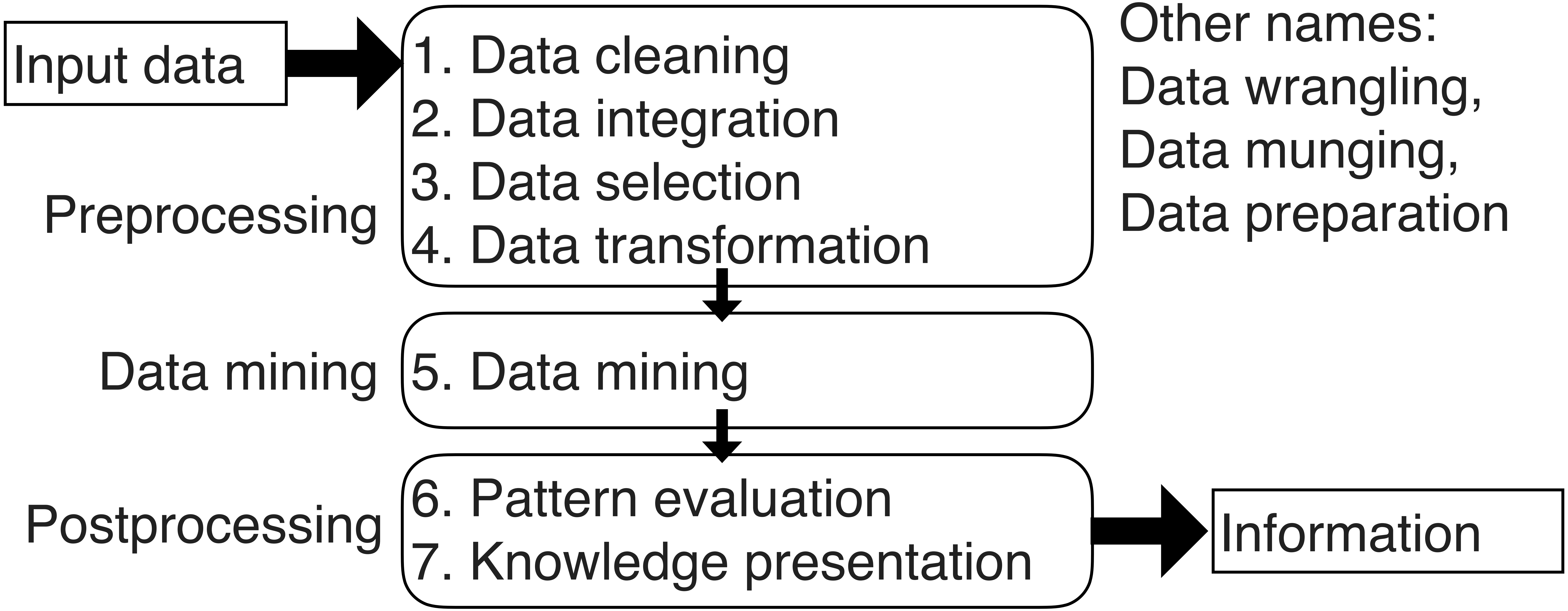
3) Inferential statistics
(Hypothesis testing)



Data mining is a data analysis technique focusing on prediction

There are many steps in data mining/analysis

Data Mining is short for Knowledge Discovery from Data (KDD):



Input data

Preprocessing

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation

Data mining

5. Data mining

Postprocessing

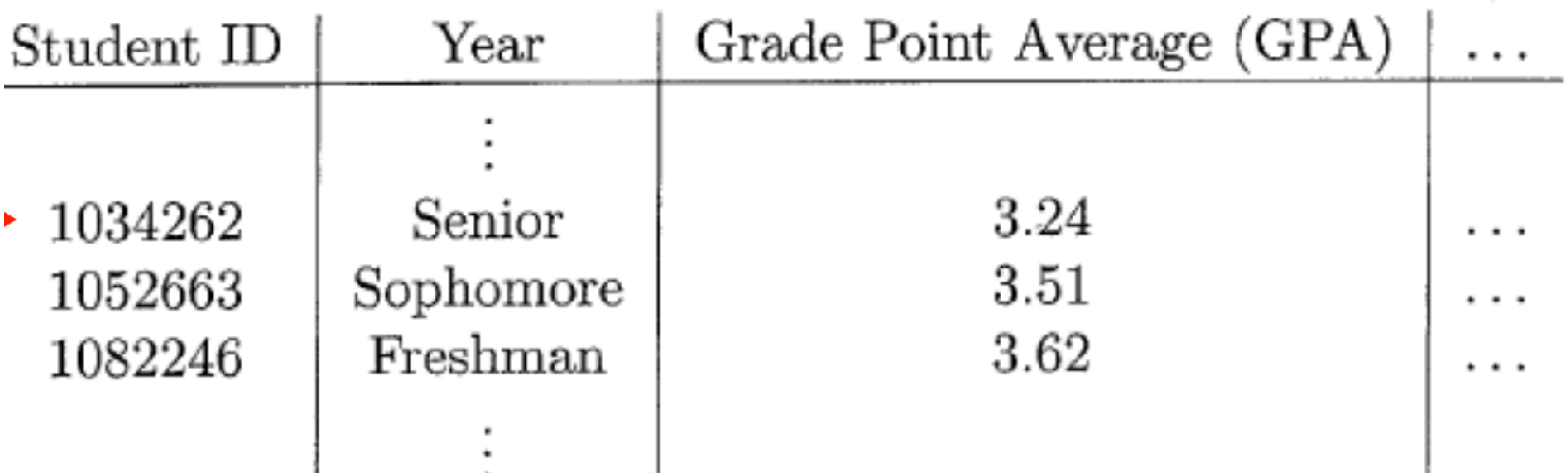
6. Pattern evaluation
7. Knowledge presentation

Information

Other names:
Data wrangling,
Data munging,
Data preparation

Data sets have objects and attributes

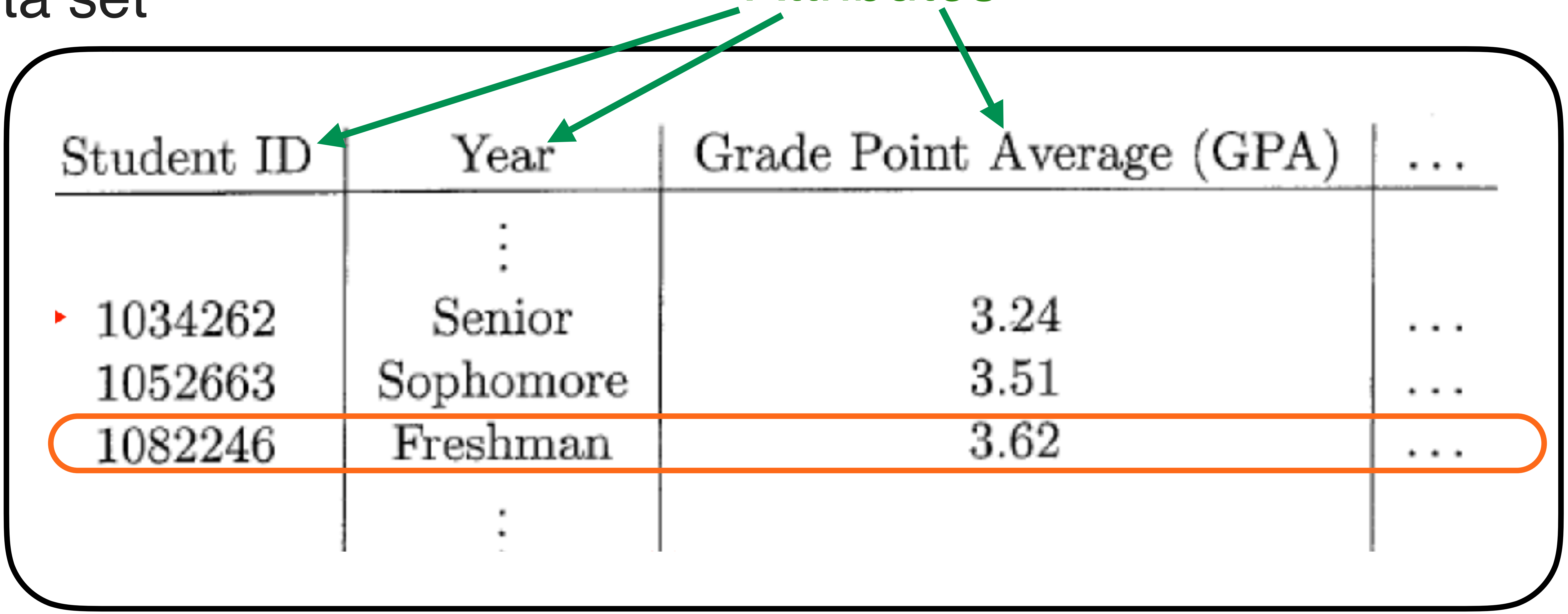
Data set



Student ID	Year	Grade Point Average (GPA)	…
	⋮		
1034262	Senior	3.24	…
1052663	Sophomore	3.51	…
1082246	Freshman	3.62	…
	⋮		

Data sets have objects and attributes

Data set



Attributes

Student ID	Year	Grade Point Average (GPA)	. . .
	⋮		
1034262	Senior	3.24	. . .
1052663	Sophomore	3.51	. . .
1082246	Freshman	3.62	. . .
	⋮		

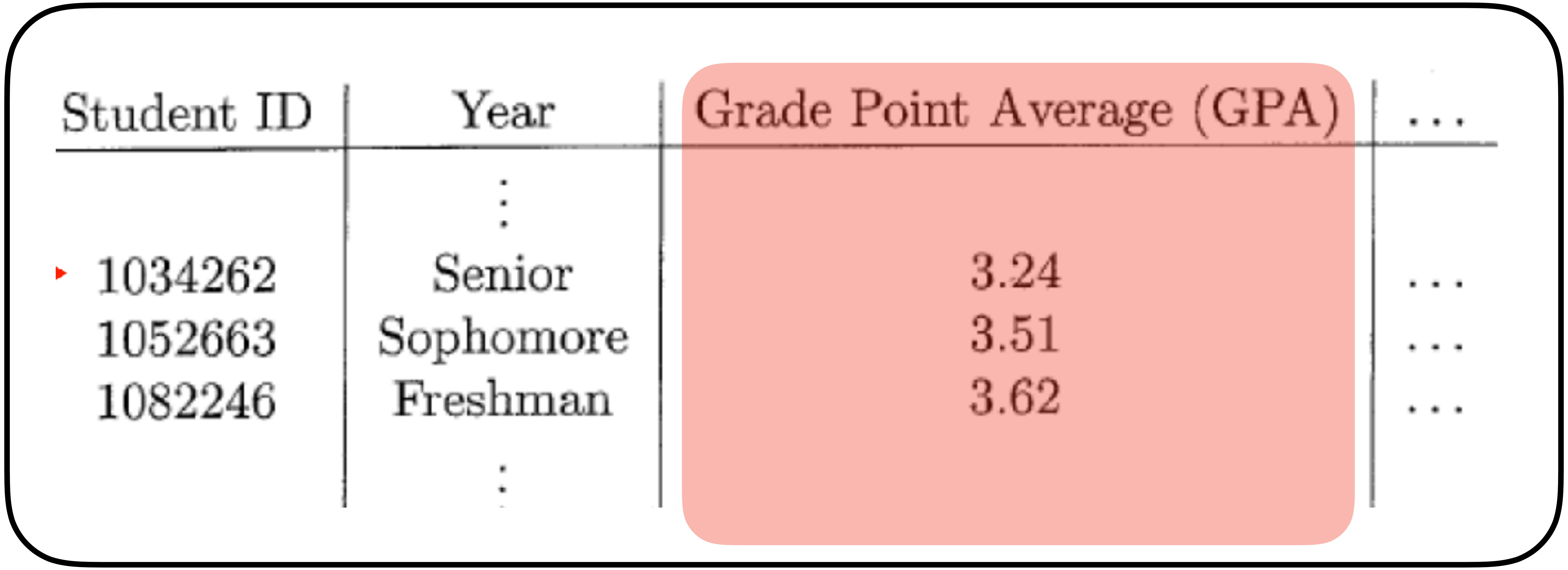
Data object

Data object = record, individual, point, event, observation, vector, entity

Attribute = field, feature, variable, dimension, characteristic

In today's class we will deal with single-variable analysis

Data set



Student ID	Year	Grade Point Average (GPA)	…
	⋮		
1034262	Senior	3.24	…
1052663	Sophomore	3.51	…
1082246	Freshman	3.62	…
	⋮		

Data object = record, individual, point, event, observation, vector, entity

Attribute = field, feature, variable, dimension, characteristic

There are two types of variables: categorical and quantitative

Places an individual into one of several categories

Categorical variables can be nominal or ordinal

Places an individual into one of several categories

No order

Order

There are two types of variables: categorical and quantitative

Places an individual into one of several categories

Takes values for which arithmetic operations make sense

Quantitative variables can be interval or ratio

Places an individual into one of several categories

Takes values for which arithmetic operations make sense

Differences meaningful

Ratios also meaningful

https://www.mentimeter.com/s/159aa50e2159b319cdec290c4375f01e/64d2c8cf2640

Categorical

Places an individual into one of several categories

Quantitative

Takes values for which arithmetic operations make sense

Nominal

Ordinal

Interval

Ratio

Quiz results

Zip code	Street number	C°	Age
Student ID			K°
Nominal	Ordinal	Interval	Ratio

Jupyter

Outliers can be a sign for low data quality

Outliers (anomalous objects or values):

1) Data objects that have characteristics different from most others, or
2) Values of an attribute that are unusual

Outliers can be a sign for low data quality

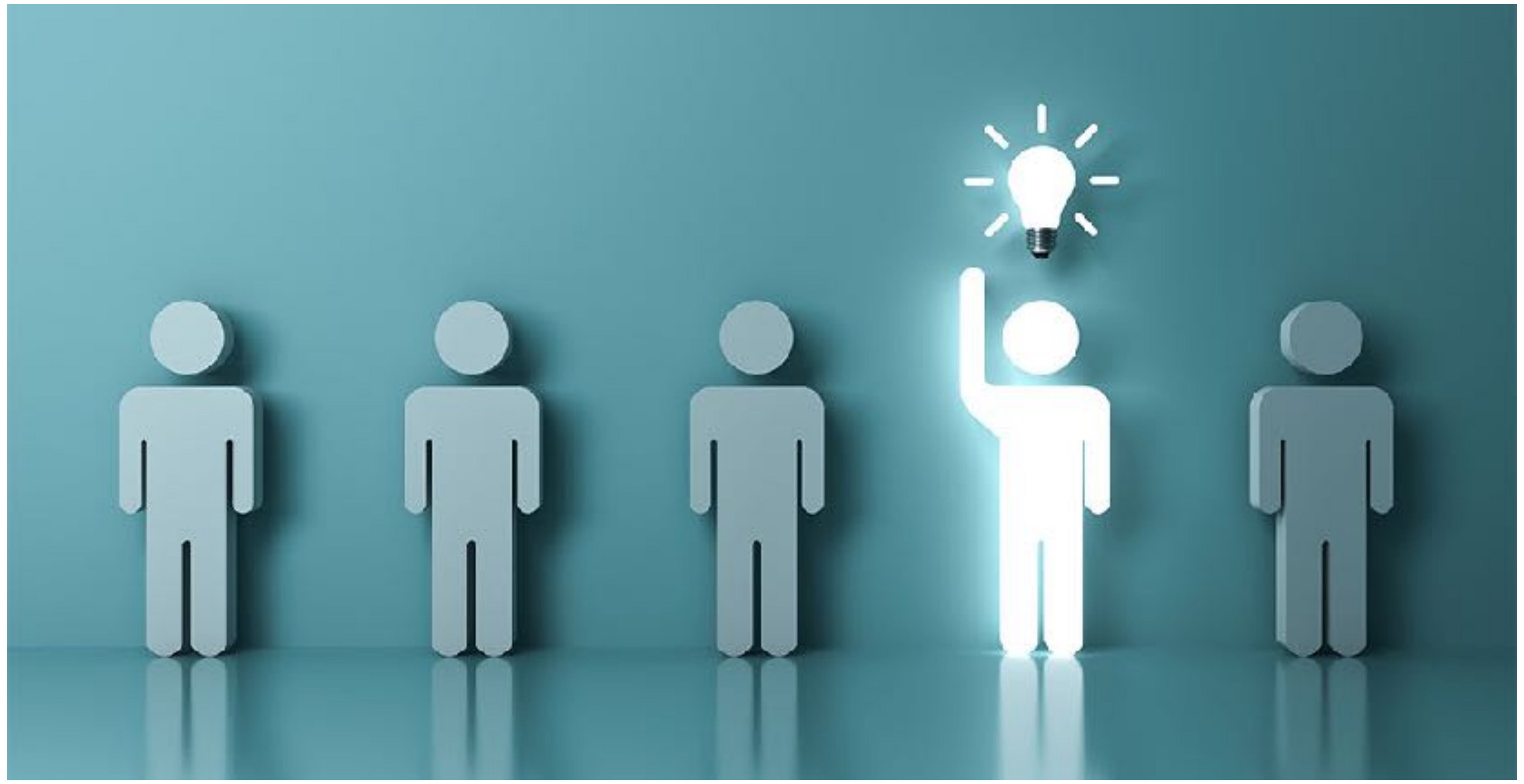
Outliers (anomalous objects or values):

1) Data objects that have characteristics different from most others, or
2) Values of an attribute that are unusual

This is not just noise! An outlier is an event that is suspected of not being generated by the same mechanisms as the rest of the data.

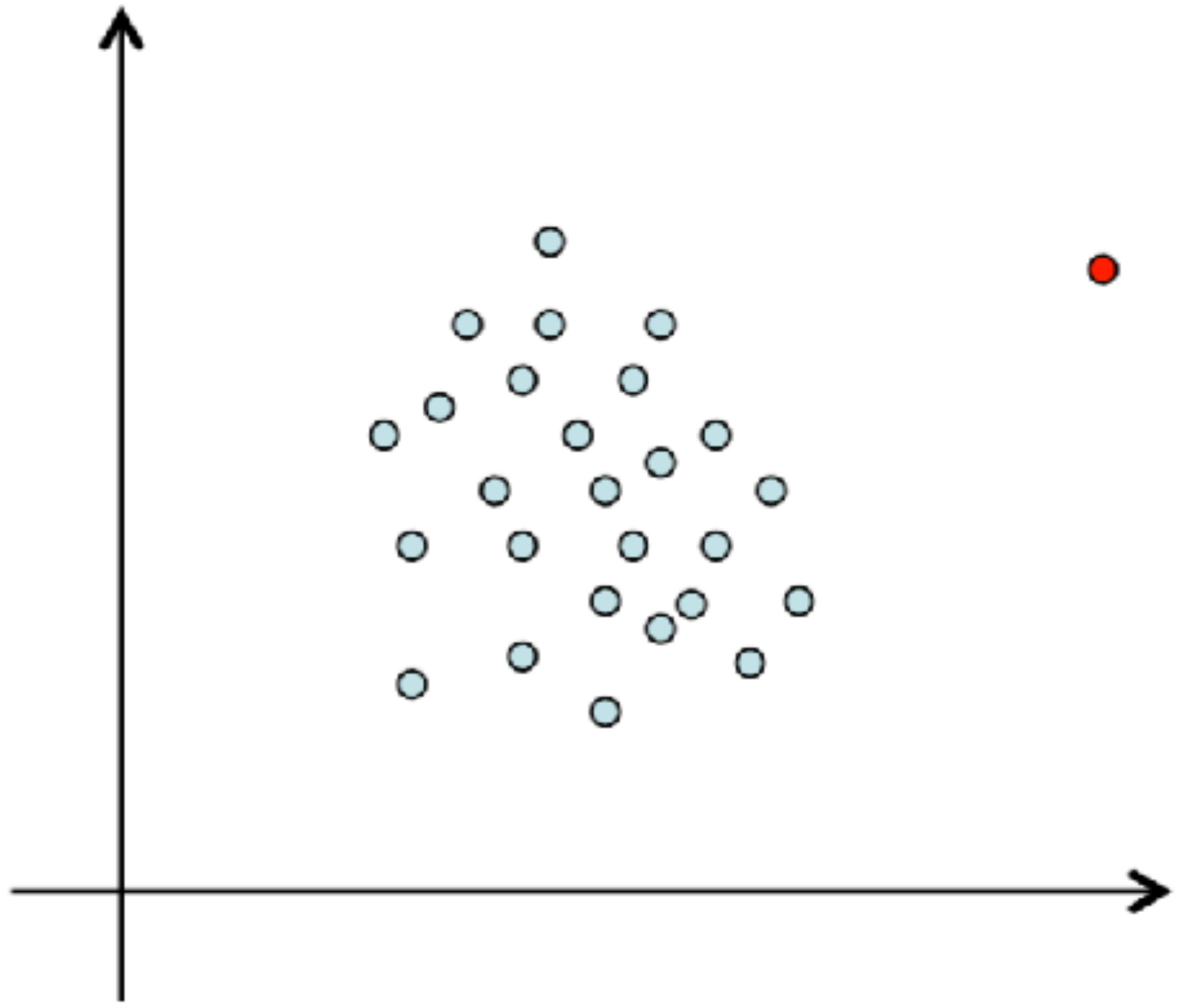
Outliers can be legitimate, interesting objects

Fraud

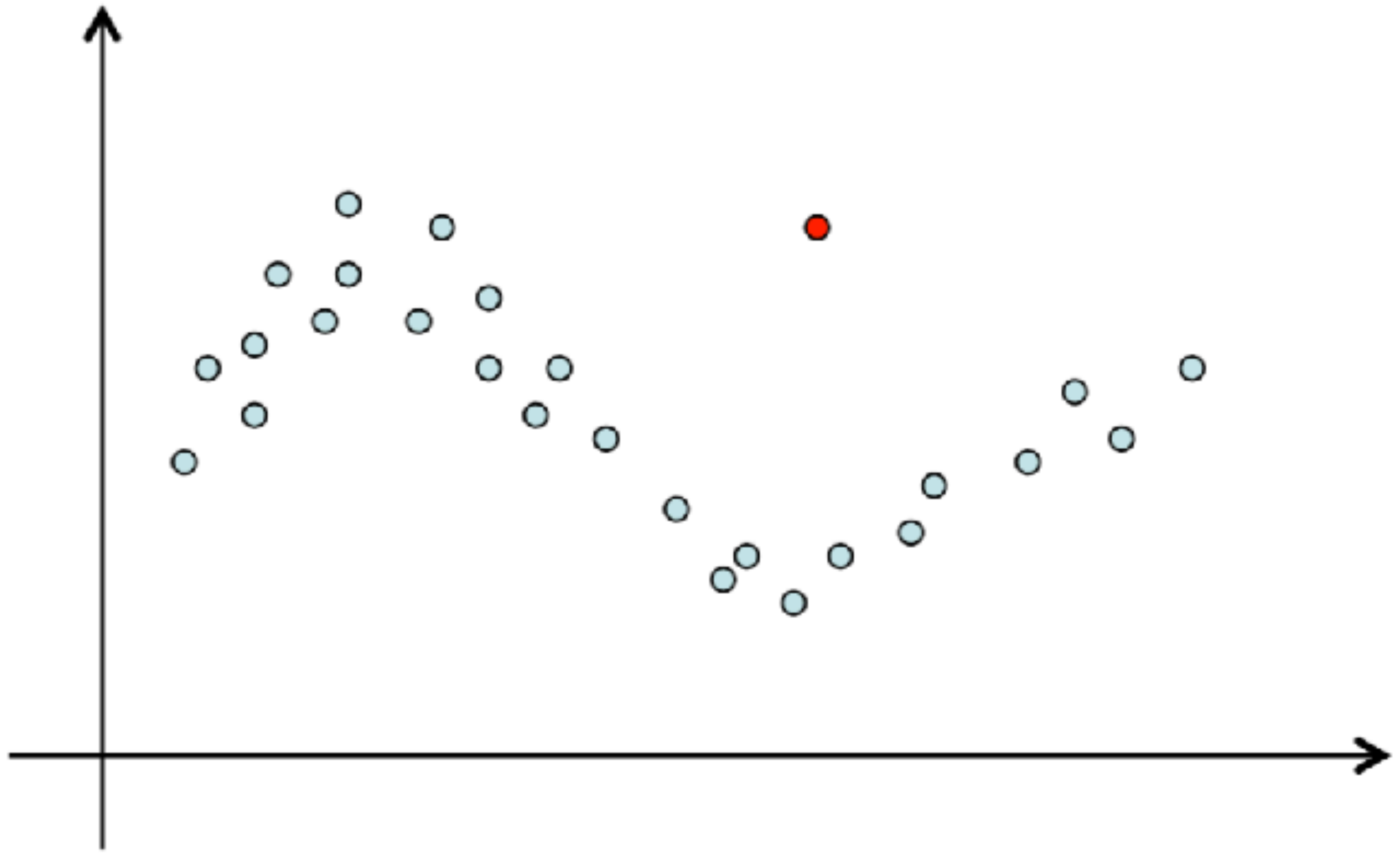


Innovation

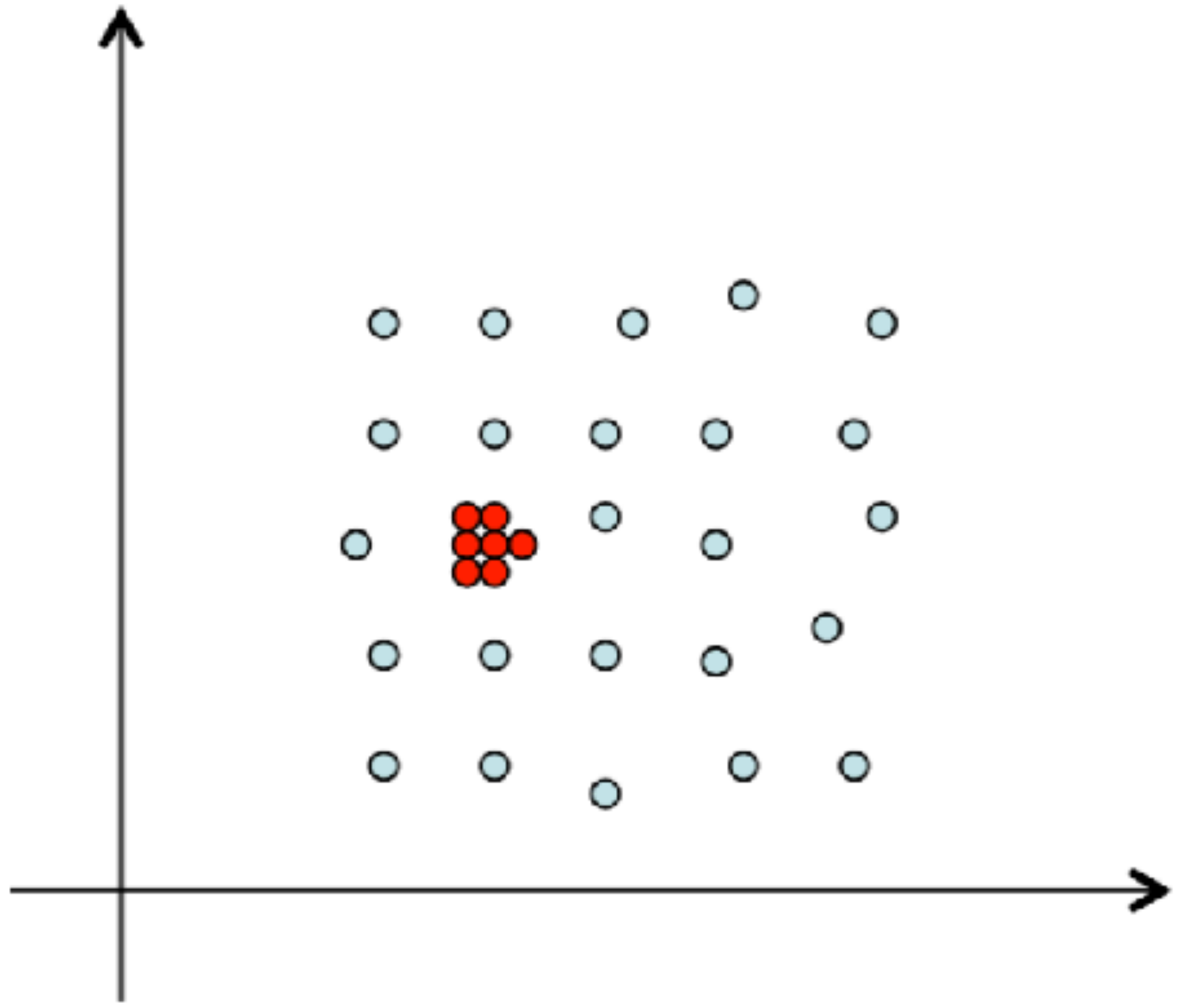
There are different kinds of outliers



Global outliers

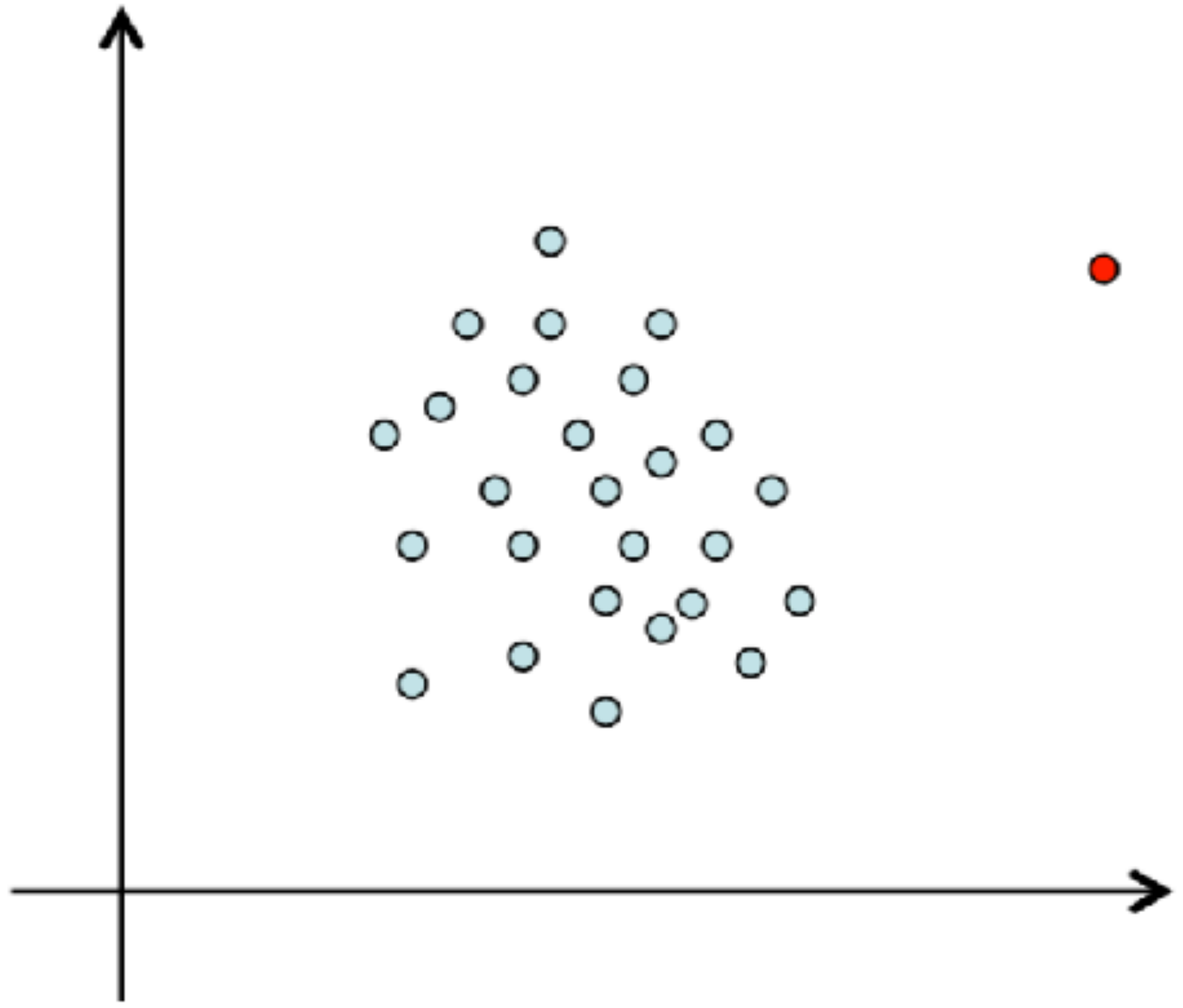


Contextual outliers



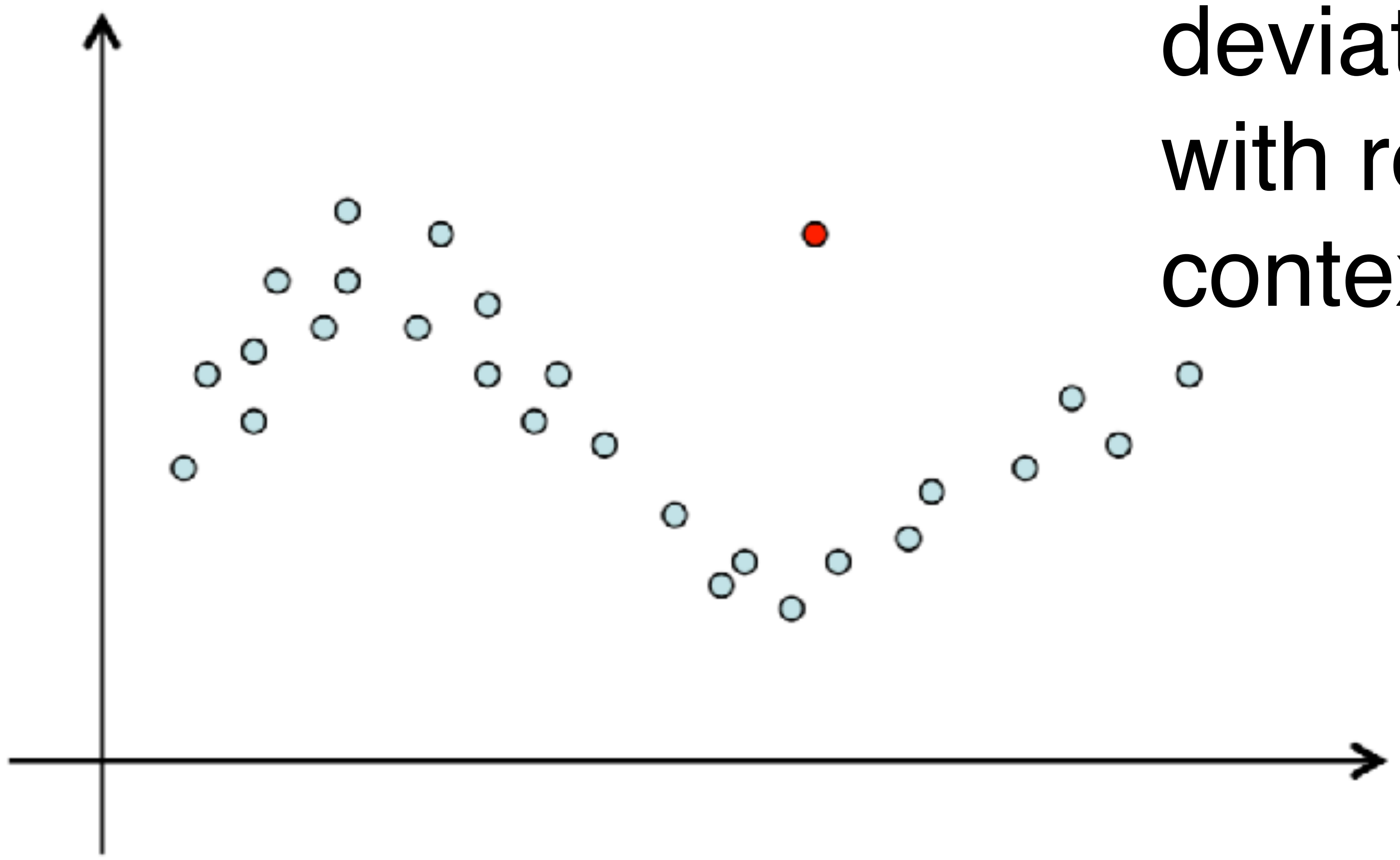
Collective outliers

There are different kinds of outliers



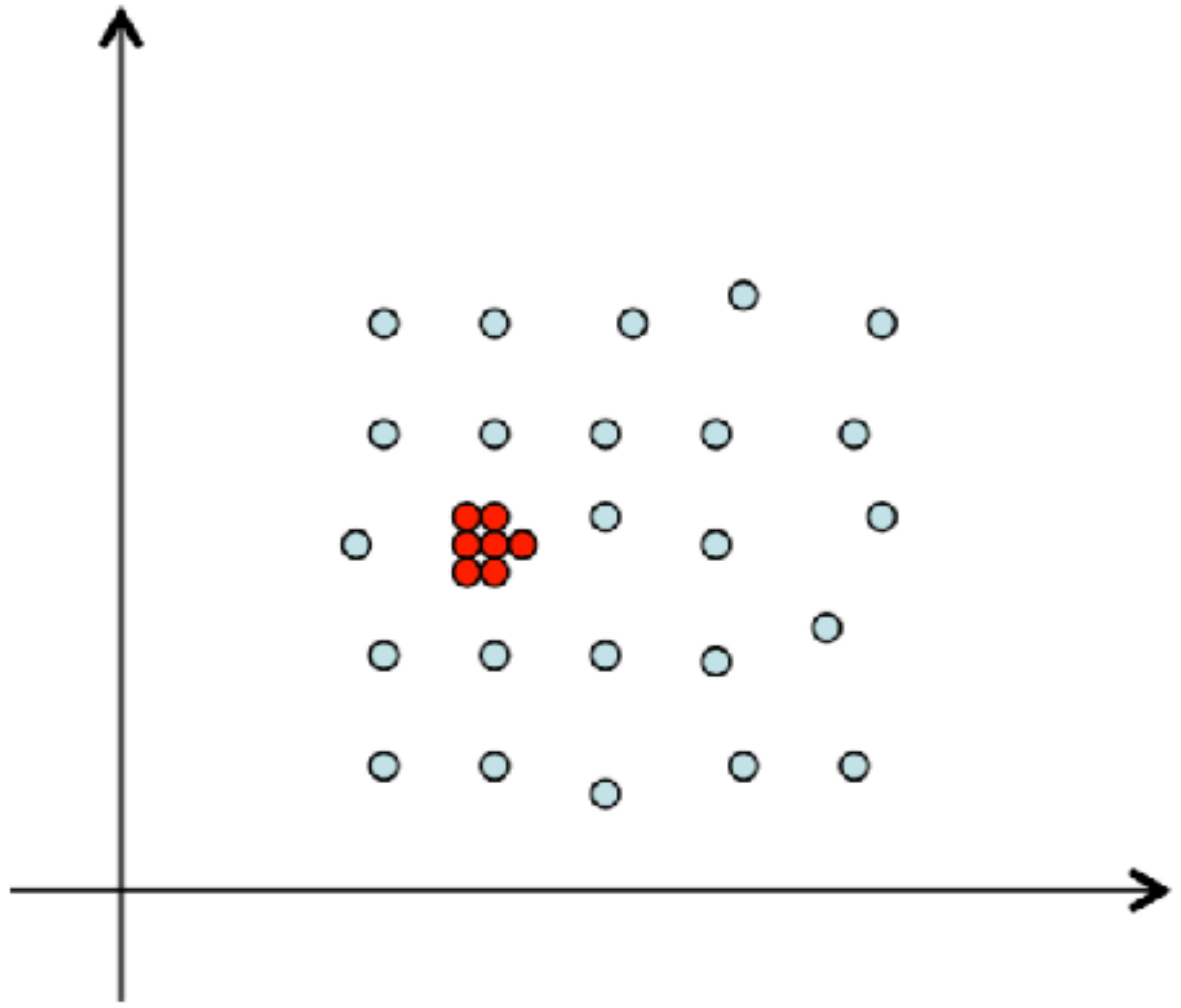
Global outliers

deviates significantly from the rest of the data. Also called: point anomaly



Contextual outliers

deviates significantly with respect to a given context of the object

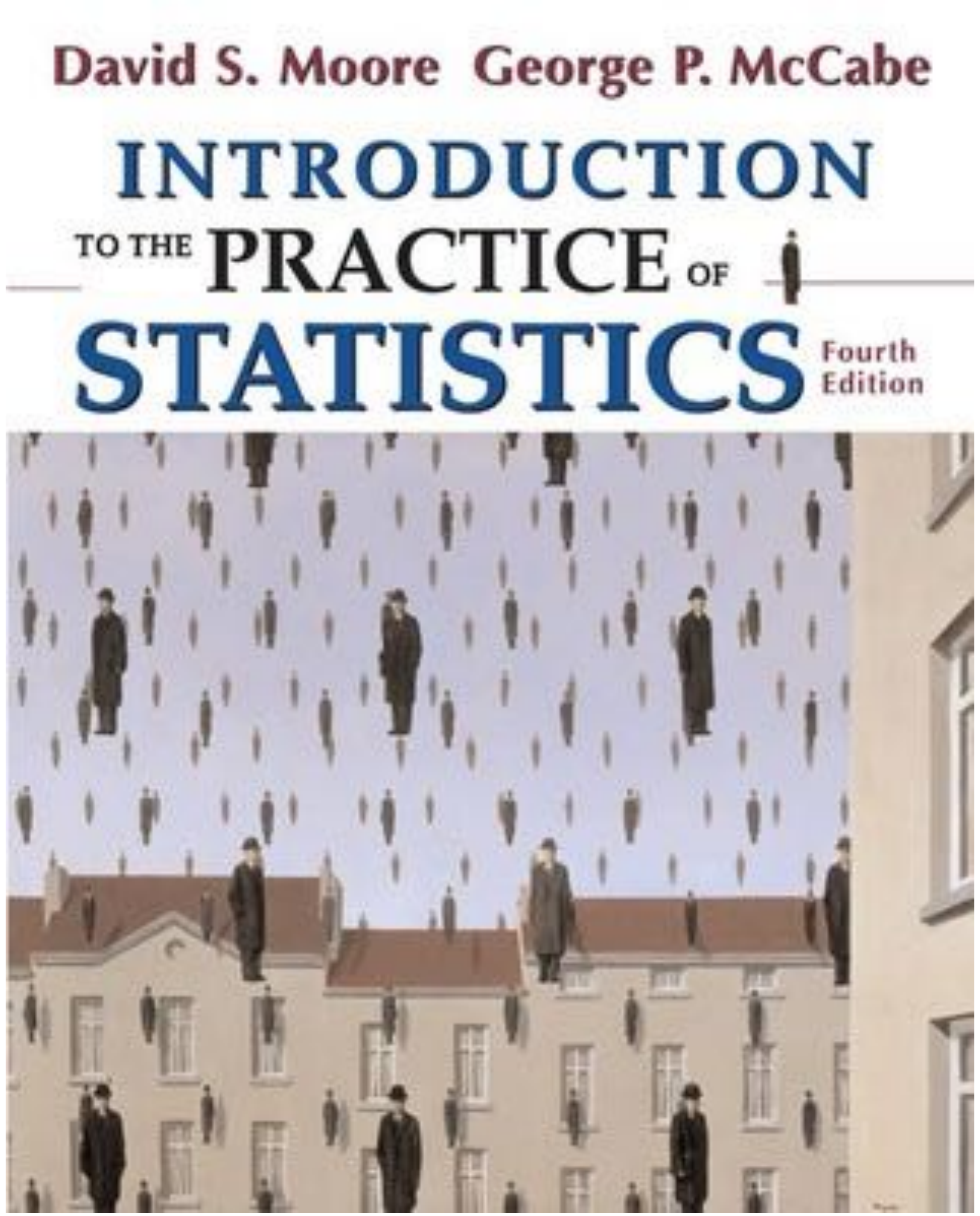


Collective outliers

a subset of data objects that as a group deviate significantly from the typical behavior of the entire data set.

An individual object of these collective outliers might not be an outlier itself.

Sources and further materials for today's class


David S. Moore George P. McCabe
INTRODUCTION TO THE PRACTICE OF STATISTICS
Fourth Edition

Chapter 1