Data Valorization:
Introduction and Data Description

Lionel Fillatre

fillatre@unice.fr

Jalon website

- Class website

https://lms.univ-cotedazur.fr/course/view.php?id=4098

- Password for the course registration: **wmg578GTA**

- **You must register yourself!**

Teachers

- Lecture (english): M. Lionel FILLATRE

- Labs EIT-Digital (english): M. Lionel FILLATRE
  (today in E-105)
- Labs MAM4-IMAFA (french): M. Cyprien GILET
- Labs MAM4-SD (french): M. Ayoud BADIA
- Labs SI4 (french): M. Cyprien GILET

Topics

- Introduction
- Examination
- What is R?
- What is Data?
- Data Sets
- Conclusion

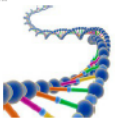
1 Introduction

Big Data Era

- ~1 trillion webpages
(http://googleblog.blogspot.dk/2008/07/we-knew-web-was-big.html)

- One hour of video is uploaded to youtube every second resulting in 10 years of content every day
(source: youtube)

- We have sequenced more than 1000 peoples genome of $3.8 \cdot 10^9$ base pairs
(source: K. P. Murphy "Machine Learning")



- Walmart handles more than 1 mio. transactions per hour and has databases containing more than $2.5 \cdot 10^{15}$ bytes of information
(source: K. P. Murphy "Machine Learning")

- Each night the worlds astronomy laboratories store high-resolution of the night sky of around a terabyte (10^{12})
(source: Stephen Marsland "Machine Learning An Algorithmic Perspective")

- In total, the four main detectors at the Large Hadron Collider (LHC) produced 13 petabytes (10^{15}) of data in 2010
(source: wikipedia "Big Data")

- Facebook handles 40 billion photos from its user base.
(source: wikipedia "Big Data")

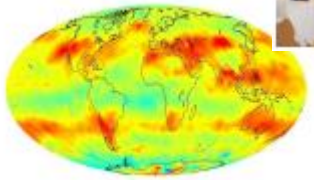
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
(source: wikipedia "Big Data")

Applications

- **Chemistry**
  - Spectrometry, Chemical sensors
- **Audio processing**
  - Spoken digit classification, Music genre classification
- **Image processing**
  - Hand-written digit recognition, Image tagging and classification, Number plate recognition
- **Informatics**
  - Collaborative filtering, Text corpus Analysis, Spam filters, Computer games
- **Biomedical**
  - Micro-array gene analysis, Medical Imaging
- **Financial data mining**
  - Market predictions
- **Climate data**
  - Weather forecast

NETFLIX

amazon.com



Deluge of information

Every day, we create 2.5 quintillion (10^{18}) bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

Source: http://www-01.ibm.com/software/data/bigdata/

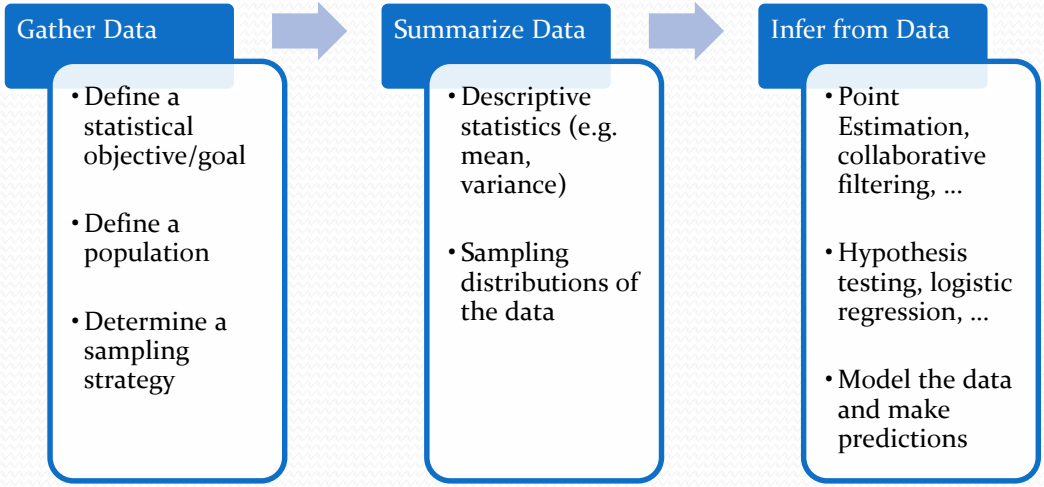
"If data had mass, the earth would be a black hole"

Stephen Marsland

"We are drowning in information and starving for knowledge"

John Naisbitt

Statistical Inference in a Nutshell



Topics Covered

3 parts: basics in data manipulation with R (1-4), basics in statistics inference (5-9), applications (10-11)

1. Introduction and Data Description 06/02/2019
2. Data Gathering and Sampling 13/02/2019 => Kaggle team registration
3. Data Visualization 27/02/2019 => 1st written exam (no official lecture)
4. Shiny Application 06/03/2019
5. Point Estimation 13/03/2019
6. Logistic Regression 20/03/2019
7. Hypothesis Testing 27/03/2019 => 1st Kaggle delivery
8. Naïve Bayes Test 03/04/2019
9. Correspondence Analysis 10/04/2019
10. Recommendation System 24/04/2019 => 2nd written exam (no official lecture)
11. Reinforcement Learning 15/05/2019 => 2nd Kaggle delivery

Prerequisites

- "Fluency" with basic probability and analysis
  - Random variables
  - Probability distributions, joint distributions, conditional probability
  - Independence/Correlation/Covariance
  - Law of Large Numbers
  - Central Limit Theorem
- Multivariable calculus together with linear algebra are required.
- R is not required.

2 Examination

Examination

- Written examination
  - 1 hour per exam during the lecture (no lecture the weeks of the exam)
  - 2 exams: 1 midterm and 1 final
- Kaggle Challenge: 2 deliveries in R
- Final grade based on an overall assessment of the deliveries in R and written exams (**4 grades, 25% each one**).

Written Exams

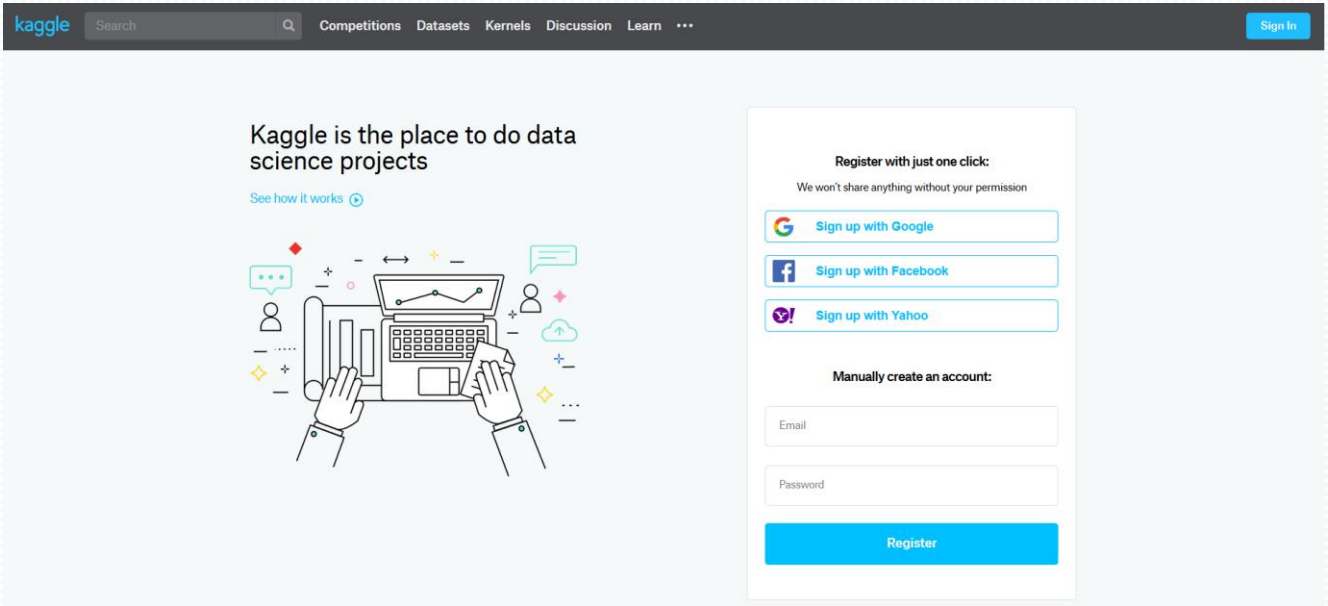
- First exam
  - Essentially based on the refresher training (related to the probability theory and the statistical theory)
  - At least one exercise related to « Data Valorization »
  - Rather simple if you work carefully the refresher training

- Second exam
  - All the exercises related to « Data Valorization »

Kaggle

- Kaggle is a platform for predictive modelling and analytics competitions
- Companies and researchers post their data
- Statisticians, data miners, data scientists (and others) from all over the world compete to produce the best models.
- Website: https://www.kaggle.com/

Kaggle website



Kaggle website

- **Competitions**
  - The competition host prepares the data and a description of the problem
- **Datasets**
  - With or without competition
- **Kernels**
  - Kernels contain both the code needed for an analysis, and the analysis itself. It's the core of a work, what it needs to make it reproducible, to make it grow, and to invite collaboration.
- **Discussion**: forum of discussions
- **Jobs**: Hiring? Seeking?
- **Learn**: learn the basics to confidently start a new career or upgrade your skills.
- **Blog**: official blog of Kaggle.com
- **User rankings**: ranking of Kaggle users
- **Tags**: to find pages associated to a specific tags
- **Host a competition**: Kaggle can help you solve difficult problems, recruit strong teams, and amplify the power of the data science talent.

Kaggle Challenge

- Work in group of 4 students maximum (3 students minimum)
- You have to choose one dataset (each team must have its own dataset; no duplication; Titanic dataset is not allowed)
- Choose the challenge according to your skills (decide the balance between mathematics and computer science)
- Team registration:
  - Fill in the shared file on the class website: name of the students and name of the challenge
  - Once a given challenge is written in the shared file, the remaining teams must choose an other challenge

Kaggle Challenge

- You have to define the question you aim to answer (classification, dimension reduction, regression, etc.)
- You have to provide numerical and theoretical justification of your analysis
- I encourage you to reuse an existing Kaggle Kernel
- You can write, run, and view best practice code and visualizations of this dataset on Kaggle Kernels.
- **You must exploit the theoretical tools and practical methods presented in this course!**

Deliveries of the challenge

- 1st delivery:
  - Recorded oral presentation (10 minutes in video) of the chosen challenge (business goal, technical goal, data description, statistical analysis of data, data preprocessing in R to extract relevant information, brief description of the future analysis which will be detailed in the 2nd delivery, etc.)

- 2nd delivery:
  - Recorded oral presentation (10 minutes in video) of the chosen Kernel in R (you can create your own notebook by merging several notebooks)
  - You must promote an advanced analysis (see next slide)
  - Present the method with **theoretical explanations** and the results
  - Run the notebook with explanations
  - The final notebook must be uploaded on the class website. It must be easy to follow (well commented).

- For each presentation, all team members must talk and you can use a slideshow.
- For each video, record the video and post it somewhere (YouTube, etc.). Send an email with the URL of the video to your labs professor.

Management is the key

- Nothing to invent, just discover and UNDERSTAND!
- Many tasks already identified:
  - R to discover
  - Dataset to understand
  - Data to load and to analyze
  - Data to preprocess
  - Choose only one advanced data analysis method and use it correctly:
    - Logistic regression
    - Random forest
    - Deep learning
    - …
  - Results to produce and to analyze
  - Presentation to prepare and to record

Important dates

- Kaggle: three deadlines at 21h00 (-1 per day late)
- Written exams: possible change in case of unexpected issues

1. Introduction and Data Description 06/02/2019
2. Data Gathering and Sampling 13/02/2019 => Kaggle team registration
3. Data Visualization 27/02/2019 => 1st written exam (no official lecture)
4. Shiny Application 06/03/2019
5. Point Estimation 13/03/2019
6. Logistic Regression 20/03/2019
7. Hypothesis Testing 27/03/2019 => 1st Kaggle delivery
8. Naïve Bayes Test 03/04/2019
9. Correspondence Analysis 10/04/2019
10. Recommendation System 24/04/2019 => 2nd written exam (no official lecture)
11. Reinforcement Learning 15/05/2019 => 2nd Kaggle delivery

3 What is R?

Useful programming languages in Data Science and Big Data

- SQL (1970): querying and namaging data
- Python (1991): data processing, productivity, good learning curve
- **R (1995): data analysis, oriented toward statistical analysis, more difficult to learn, free alternative to SAS, huge community**
- And others: Java, Scala, SAS, Matlab

History of R

- S: language for data analysis developed at Bell Labs circa 1976
- Licensed by *AT&T/Lucent* to *Insightful Corp*.
- R: initially written and released as an open source software by Ross Ihaka and Robert Gentleman at U Auckland during 90s (R plays on name "S")
- Since 1997: international R-core team ~15 people & 1000s of code writers and statisticians happy to share their libraries!

Open source

- Free but also much more:
  - Provides full access to algorithms and their implementation
  - Gives you the ability to fix bugs and extend software
  - Provides a forum allowing researchers to explore and expand the methods used to analyze data
  - Ensures that scientists around the world - and not just ones in rich countries - are the co-owners to the software tools needed to carry out research
  - Promotes reproducible research by providing open and accessible tools
  - Most of R is written in… R! This makes it quite easy to see what functions are actually doing.

What is it?

- R is an interpreted computer language.
  - Most user-visible functions are written in R itself, calling upon a smaller set of internal primitives.
  - It is possible to interface procedures written in C, C+, or FORTRAN languages for efficiency, and to write additional primitives.
  - System commands can be called from within R

- R is used for data manipulation, statistics, and graphics. It is made up of:
  - operators (+ - <- * %*% …) for calculations on arrays & matrices
  - large, coherent, integrated collection of functions
  - facilities for making unlimited types of publication quality graphics
  - user written functions and sets of functions (packages); 800+ contributed packages so far and growing

R for Practical Works

- We will use R with R-Studio Desktop (an interactive R development environment).
- Please visit the page https://cran.r-project.org/
- Please visit the page https://www.rstudio.com/
- Install R-Studio

R for the Labs

- You can use either **R Notebook** or **Jupyter Notebook with R**
- R Notebook is contained in RStudio
- How to use the Jupyter notebook with R?
  - If necessary, install the Jupyter Notebook. An easy way is to install Anaconda (a Python distribution) which already contains the Jupyter Notebook.
    Please visit the page https://www.anaconda.com/
  - Once Jupyter Notebook is installed, please visit the following page to run R in the Jupyter Notebook
    https://irkernel.github.io/installation/
  - Then just create a Jupyter Notebook associated to the R kernel, and fill in the notebook with R commands.

Web links

- http://www.cyclismo.org/tutorial/R/
- https://www.tutorialspoint.com/r/
- http://www.statmethods.net/index.html
- http://zoonek2.free.fr/UNIX/48_R/all.html

4 What is Data?

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object (also known as variable, field, characteristic, or feature)
- Collection of attributes describe an object (also known as record, point, case, sample, entity, or instance)

Attributes

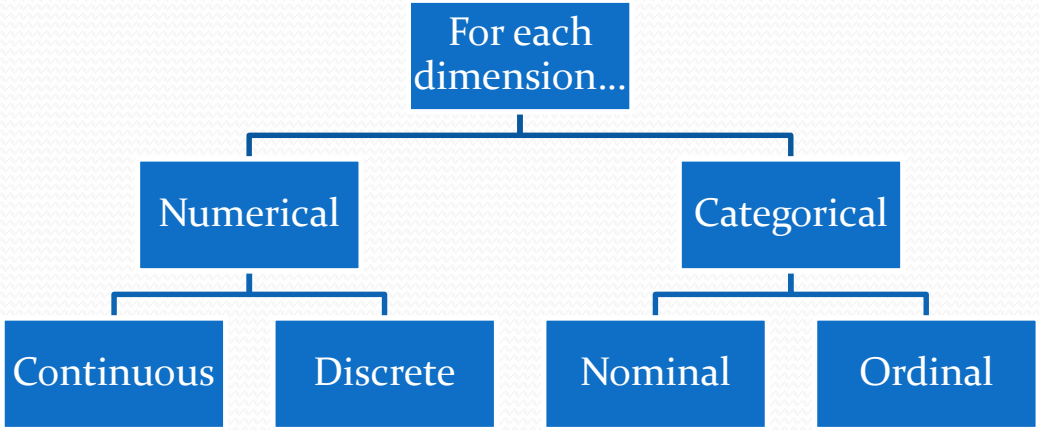
Data objects

ID	Age	Gender	Name
1	31	F	Alex
2	24	M	Ben
3	52	F	Cindy
4	35	M	Dan
5	58	M	Eric
6	46	F	Fay
7	42	M	George

Data Classification

- Data/Sample: $(X_1, \dots, X_n)$
- Dimension of X_i (i.e. the number of measurements per unit i)
  - Univariate: one measurement for unit i (height)
  - Multivariate: multiple measurements for unit i (height, weight, gender)
- For each dimension, X_i can be numerical or categorical
- Numerical variables
  - Discrete: human population, natural numbers, (0,5,10,15,20,25,etc..)
  - Continuous: height, weight
- Categorical variables
  - Nominal: categories have no ordering (gender: male/female)
  - Ordinal: categories are ordered (grade: A/B/C/D/F, rating: high/low)

Data Types


For each dimension…
Numerical
Categorical
Continuous
Discrete
Nominal
Ordinal

Summaries for numerical data

- **Center/location:** measures the "center" of the data
  - Examples: sample mean and sample median
- **Spread/Dispersion:** measures the "spread" or "fatness" of the data
  - Examples: sample variance, interquantile range
- **Order/Rank:** measures the ordering/ranking of the data
  - Examples: order statistics and sample quantiles

Summary	Type of Sample	Formula	Notes
Sample mean, $\hat{\mu}, \bar{X}$	Continuous	$\frac{1}{n}\sum_{i=1}^{n} X_i$	• Summarizes the "center" of the data • Sensitive to outliers
Sample variance, $\widehat{\sigma^2}, S^2$	Continuous	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	• Summarizes the "spread" of the data • Outliers may inflate this value
Order statistic, $X_{(i)}$	Continuous	i^{th} largest value of the sample	• Summarizes the order/rank of the data
Sample median, $X_{0.5}$	Continuous	If n is even: $\frac{\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right)}{2}$ If n is odd: $X_{\left(\frac{n}{2}+0.5\right)}$	• Summarizes the "center" of the data • Robust to outliers
Sample α quartiles, X_α $0 \le \alpha \le 1$	Continuous	If $\alpha = \frac{i}{n+1}$ for $i = 1, \ldots, n$: $X_\alpha = X_{(i)}$ Otherwise, do linear interpolation	• Summarizes the order/rank of the data • Robust to outliers
Sample Interquartile Range (Sample IQR)	Continuous	$X_{0.75} - X_{0.25}$	• Summarizes the "spread" of the data • Robust to outliers

Multivariate numerical data

- Each dimension in multivariate data is univariate and hence, we can use the numerical summaries from univariate data (e.g. sample mean, sample variance)

- However, to study two measurements and their relationship, there are numerical summaries to analyze it: **Sample Correlation** and **Sample Covariance**

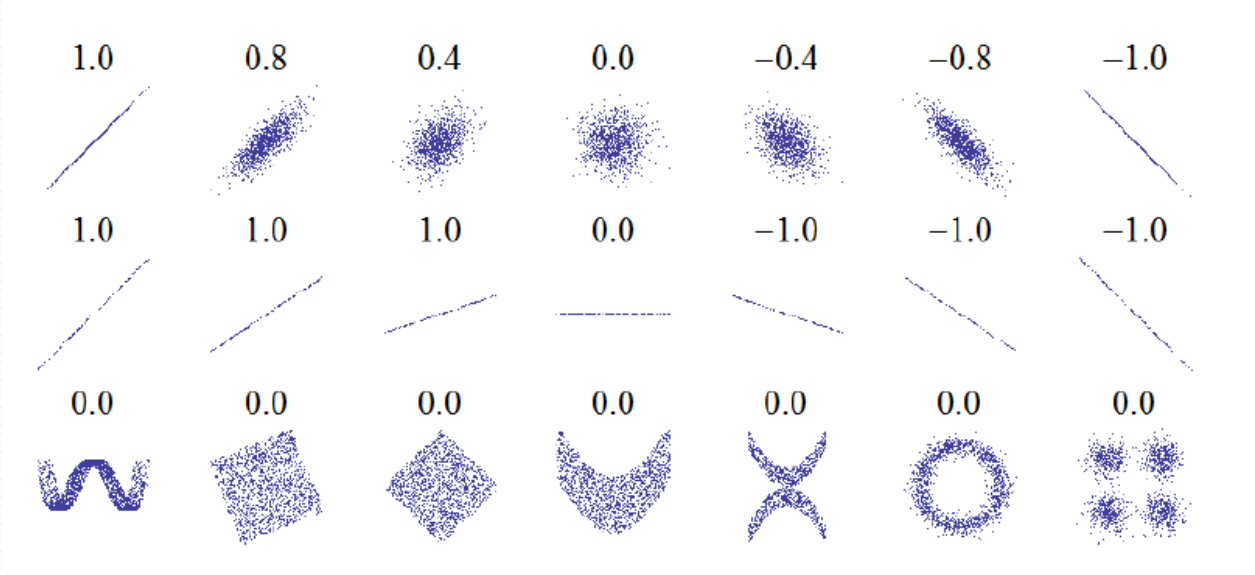
Sample Covariance and Correlation

- Measures **linear relationship** between two measurements, X_{i1} and X_{i2}, where $X_i = (X_{i1}, X_{i2})$

- Sample covariance: $\hat{\sigma}_{X_1,X_2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = \hat{\rho}\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}$

- Sample correlation: $\hat{\rho} = \frac{\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{(n-1)\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}$
  - $-1 \leq \hat{\rho} \leq 1$
  - Sign indicates proportional (positive) or inversely proportional (negative) relationship
  - If X_{i1} and X_{i2} have a maximum or minimum sample correlation ($\hat{\rho} = 1$ or -1), then $X_{i2} = aX_{i1} + b$

Correlation



Summaries for categorical data

- **Frequency/Counts:** how frequent is one category
- Generally use tables to count the frequency or proportions from the total
- Example: class composition

	Undergrad	Graduate	Staff
Counts	17	1	2
Proportions	0.85	0.05	0.1

5 Data Sets

Main Types of data sets

- Record data
  - Collection of data objects and their attributes
  - Representation: Table
- Ordered data
  - Ordered collection of data objects
  - Representation: Sequence
- Relational data
  - Collection of data objects and their relation
  - Representation: Graph

Record data example: Market basket data

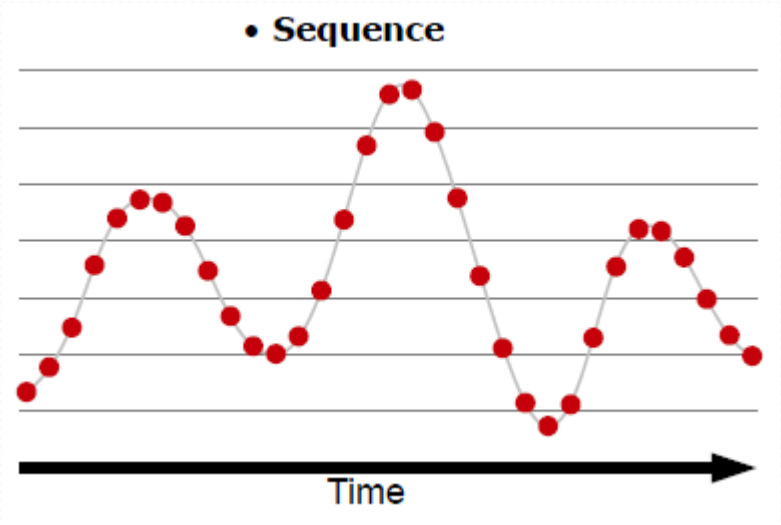
- **Transaction data table**

ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

- **Matrix**

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Ordered data example: Time series

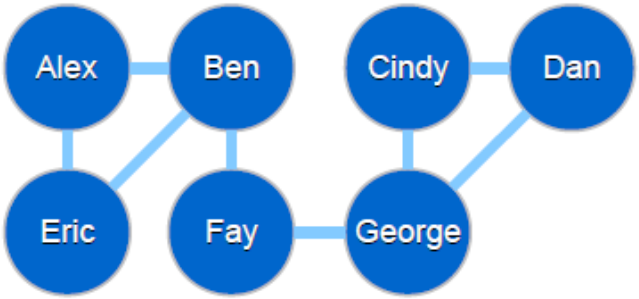


• Matrix

Time	Value
0	1.3
1	1.8
2	2.5
3	3.6
4	4.4
5	4.7
6	4.6
7	4.3
8	2.4
9	2.1
10	2.0
11	2.3
12	3.1

Relational data example: Who knows who?

• Graph



• Matrix

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	0	0	1	1	0
C	0	0	0	1	0	0	1
D	0	0	1	0	0	0	1
E	1	1	0	0	0	0	0
F	0	1	0	0	0	0	1
G	0	0	1	1	0	1	0

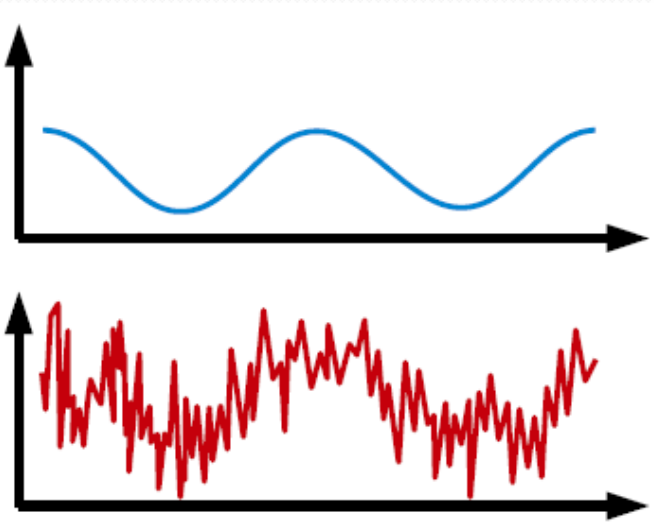
Data quality

- Data is of high quality if they
  - Are fit for their intended use
  - Correctly represent the phenomena they correspond to

- Examples of quality problems
  - Noise
  - Outliers
  - Missing values

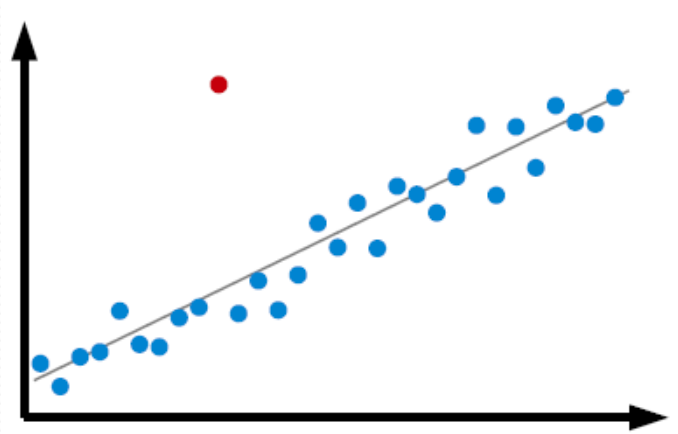
Noise

- Definition
  - Unwanted perturbation to a signal
  - Unwanted data
- Reasons for noise
  - Fundamental limits in measurement accuracy
  - Interference from other signals
  - Measurement of attributes not related to the data modeling task
- Handling noise
  - Exclude noisy attributes
  - Remove noise by filtering
  - Include a model of the noise



Outliers

- Definition
  - Data objects which are significantly different from most others
- Reasons for outliers
  - Measurement error
  - Natural property of data
- Handling outliers
  - Identify outliers
  - Exclude anomalous outliers
  - Model the outliers



Missing values

- Definition
  - No value is stored for an attribute in a data object
- Reasons for missing values
  - Information is not collected
    - People decline to give their age
  - Attribute is not applicable
    - Annual income is not applicable to children
- Handling missing values
  - Eliminate data objects
  - Estimate missing values
  - Ignore the missing value in analysis
  - Replace with an average value

ID	Age	Gender	Name
1	31	F	Alex
2	(?)	M	Ben
3	52	F	Cindy
4	35	(?)	Dan
5	(?)	M	Eric
6	(?)	F	Fay
7	42	M	(?)

6 Conclusion

Conclusion

- Data must be clean or cleaned
- Statistics is necessary to clean the data, then to analyze it
- Analysis depends on data types
- Data summaries are very useful to clean the data and verify the quality of observations