# Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@ce.yildiz.edu.tr

https://www.ce.yildiz.edu.tr/personal/zeyneb

# Course Details

- Course Code: BLM3590
- Course Name: Statistical Data Analysis
- Credit: 3
- Nature of the course: Lecture
- Course web page:

https://www.ce.yildiz.edu.tr/personal/zeyneb/file/34457/BLM3590_IstatistikselVeriAnalizi_20182

- Instructor: Assist. Prof. Dr. Zeyneb KURT

Email: zeyneb@ce.yildiz.edu.tr

# Rules of the Conduct

- You are responsible for checking the class web page often for announcements.
  - https://www.ce.yildiz.edu.tr/personal/zeyneb
- Academic dishonesty and cheating
  - will not be tolerated
  - will be dealt with according to university rules and regulations
    - http://www.yok.gov.tr/content/view/475/
    - Presenting any work that does not belong to you is also considered academic dishonesty.

# Attendance Policy

- If you are late to class >10 minutes, you will not be allowed to enter the classroom.

- The requirement for attendance is **70%**
  - **The students, who fail to fulfill the attendance requirement, will be excluded from the final exams and the grade of F0 will be given.**

# Assesment

| | | |
|---|---|---|
| • Quiz | : | 10% |
| • Midterm(s) | : | 25% |
| • Homework(s) | : | 20% |
| • Final | : | 40% |
| • Attendance & participation | : | 05% |

(The requirement for attendance is 70%)

# Objective

- Overall
  - Reinforce your understanding of statistical data analysis
- Specific
  - Concepts of data analysis
  - Some data analysis techniques
  - Some tips for data analysis
- Try to cover every bit and pieces of statistical data analysis techniques
- Suggested book: Ajit C. Tamhane & Dorothy D. Dunlop. Prentice Hall, 1999.

# Course Subjects

| Week | Subjects |
|---|---|
| **1** | Introduction: Statistical Methods in the Context of Scientific Studies. Sampling. Observational Studies and Experiments. Data Exploration and Analysis. Statistical Inference. Computation Using R. |
| **2** | Data Exploration: Data Visualization and Summary Statistics. Variable Types. Exploring Categorical Variables. Exploring Numerical Variables. Data Preprocessing. |
| **3** | Exploring Relationships: Visualizing and Summarizing Relationships between Variables. Relationships between Two Numerical Random Variables. Relationships between Categorical Variables. Relationships between Numerical and Categorical Variables. |
| **4** | Probability: Probability as a Measure of Uncertainty. Complement, Union, and Intersection. Disjoint Events. Conditional Probabilities. Independent Events. Bayes' Theorem. |
| **5** | Random Variables and Probability Distributions: Random Variables. Probability Distributions. Cumulative Distribution Function and Quantiles. |
| **6** | Estimation: Parameter Estimation. Point Estimation. Sampling Distribution. Confidence Intervals. Margin of Error. |
| **7** | Hypothesis Testing: Hypothesis Testing for the Population Mean. Statistical Significance. Hypothesis Testing using t-tests. Hypothesis Testing for Population Proportion. |
| **8** | Statistical Inference for the Relationship between Two Variables: Relationship between a Numerical Variable and a Binary Variable. Inference about the Relationship between Two Binary Variables. Inference Regarding the Linear Relationship between Two Numerical Variables. |
| **9** | Midterm 1 |
| **10** | Analysis of Variance (ANOVA): Introduction. The Assumptions of ANOVA. |
| **11** | Analysis of Categorical Variables: Pearson's $\chi2$ Test for One Categorical Variable. Pearson's $\chi2$ Test of Independence. Contingency Tables. |
| **12** | Regression Analysis: Linear Regression Models with One Binary Explanatory Variable. Statistical Inference Using Simple Linear Regression Models. Linear Regression Models with One Numerical Explanatory Variable. Model Assumptions and Diagnostics. Multiple Linear Regression. |
| **13** | Clustering: K-means Clustering. Hierarchical Clustering. Standardizing Variables Before Clustering. |
| **14** | Bayesian Analysis: Introduction. Prior &Posterior Probabilities. Bayesian Inference. Estimation. Hypothesis Testing. |

# Data analysis – "The Concept"

- Approach to de-synthesizing data, informational, and/or factual elements to answer research questions
- Method of putting together facts and figures to solve research problems
- Systematic process of utilizing data to address research questions
- Breaking down research issues through utilizing controlled data and factual information

# Categories of data analysis

- Narrative (e.g. laws, arts)
- Descriptive (e.g. social sciences)
- Statistical/mathematical (pure/applied sciences)
- Audio-Optical (e.g. telecommunication)
- Others

- Most research analyses adopt the first three
- The second and third are most popular in pure, applied, and social sciences

# Statistical Methods

- Something to do with "statistics"
  - Statistics
    - meaningful quantities about a sample of objects, things, persons, events, phenomena, etc.
    - Widely used in many fields (social sciences, engineering, etc.)
    - Simple to complex issues. E.g.
      - correlation
      - anova
      - regression
      - econometric modelling
- Two main categories:
  - Descriptive statistics
  - Inferential statistics

# Descriptive statistics

- Use sample information to explain/make abstraction of samples "phenomena"
- Common "phenomena":
  - Association (e.g. correlation, mutual information)
  - Tendency (left-skew, right-skew)
  - Causal relationship (e.g. if X, then, Y)
  - Trend, pattern, dispersion, range
- Used in non-parametric analysis
  - e.g. chi-square, t-test, 2-way anova)

# Inferential statistics

- Using sample statistics to infer some "phenomena" of population parameters
- Common "phenomena": cause-and-effect
- One-way relationship

  $Y = f(X)$
- Multi-directional relationship

  $Y1 = f(Y2, X, e1)$

  $Y2 = f(Y1, Z, e2)$
- Recursive

  $Y1 = f(X, e1)$

  $Y2 = f(Y1, Z, e2)$
- Use parametric analysis

# Differences between Descriptive and Inferential statistics-1

- Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures.
- Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what the data shows.
- With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data (samples) alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Thus, we use inferential statistics to make inferences from our sample data to more general conditions; we use descriptive statistics simply to describe what's going on in our sample data.
- In a research study we may have lots of measures (e.g. mean, standard deviation). Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider the scourge of many students, the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

# Differences between Descriptive and Inferential statistics-2

- We do not have access to the whole population we are interested in investigating, but only a limited number of data instead (i.e. samples).
- For example, we might be interested in the exam marks of all students in the UK. It is not feasible to measure all exam marks of all students in the whole UK so we have to measure a smaller sample of students (e.g. 100 students), which are used to represent the larger population of all students in the UK.
- Properties of samples, such as the mean or standard deviation, are not called parameters, but statistics. Whereas properties of populations are called as parameters. Inferential statistics allow us to use these samples to make generalizations about the populations from which the samples were drawn.
- It is, therefore, important that the samples accurately represent the population. The process of achieving this is called sampling (sampling strategies are discussed in detail later on).
- Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population. The methods of inferential statistics are (1) the estimation of population parameters and (2) testing of statistical hypotheses.

# Which one to use?

- Nature of research
  - Attempts to infer, predict, find cause-and-effect, influence, relationship in the whole population or only among the samples?

# Common mistakes– “Abuse of statistics”

| Issue | Data analysis techniques | |
|---|---|---|
| | Example of abuse | Correct technique |
| Measure the “influence” of a variable on another | Using partial *correlation* (e.g. Spearman coeff.) | Using a regression parameter |
| Finding the “relationship” between one variable with another | Multi-dimensional scaling, Likert scaling | Simple regression coefficient |
| To evaluate whether a model fits data better than the other | Using $R^2$ | Many – a.o.t. Box-Cox $\chi^2$ test for model equivalence |
| To evaluate accuracy of “prediction” | Using $R^2$ and/or F-value of a model | Hold-out sample’s mean absolute percentage error |
| “Compare” whether a group is different from another | Multi-dimensional scaling, Likert scaling | Many – a.o.t. two-way anova, $\chi^2$, Z test |
| To determine whether a group of factors “significantly influence” the observed phenomenon | Multi-dimensional scaling, Likert scaling | Many – a.o.t. manova, regression |

# How to avoid mistakes - Useful tips

- Crystalize the research problem
- Read literature on data analysis techniques
- Evaluate various techniques that can do similar things w.r.t. to research problem
- Know what a technique does and what it doesn't
- Consult people, esp. supervisor
- Pilot-run the data and evaluate results

# Principles of analysis…

- Goal of an analysis:
  - To explain cause-and-effect phenomena
  - To relate research with real-world event
  - To predict/forecast the real-world phenomena based on research
  - Finding answers to a particular problem
  - Making conclusions about real-world event based on the problem
  - Learning a lesson from the problem

# …Principles of analysis…

- Data cannot talk
- An analysis contains some aspects of scientific reasoning/argument:
  - Define
  - Interpret
  - Evaluate
  - Illustrate
  - Discuss
  - Explain
  - Clarify
  - Compare
  - Contrast

# …Principles of analysis

- An analysis must have four elements:
  - Data/information
  - Scientific reasoning/argument
    - what? who? where? how? what happens?
  - Finding
    - what results?
  - Lesson/conclusion
    - so what? so how? therefore, …

# Principles of data analysis…

- Basic guide to data analysis:
  - Analyze, not narrate
  - Go back to research flowchart
  - Break down into research objectives and research questions
  - Identify phenomena to be investigated
  - Visualize the expected answers
  - Validate the answers with data
  - Do not tell something not supported by data

# …Principles of data analysis…

| Shoppers | Number |
| --- | --- |
| Male | |
| Old | 6 |
| Young | 4 |
| Female | |
| Old | 10 |
| Young | 15 |

- More female shoppers than male shoppers
- More young female shoppers than young male shoppers

# …Principles of data analysis

- When analyzing:
  - Be objective
  - Be accurate
  - Be true
- Separate facts and opinion
- Avoid “wrong” reasoning/argument.
  - E.g. mistakes in interpretation.

# Information Systems:

# **Fundamentals**

# Informatics

- The term informatics broadly describes the study and practice of
  - creating,
  - storing,
  - finding,
  - manipulating
  - sharing

  information.

# **Informatics** - Etymology

- In 1956 the German computer scientist Karl Steinbuch coined the word Informatik
  - [*Informatik: Automatische Informationsverarbeitung* ("Informatics: Automatic Information Processing")]
- The French term informatique was coined in 1962 by Philippe Dreyfus
  - [Dreyfus, Phillipe. L'informatique. Gestion, Paris, June 1962, pp. 240–41]
- The term was coined as a combination of information and automatic to describe the science of automating information interactions

# **Informatics** - Etymology

- The morphology—informat-ion + -ics—uses
- the accepted form for names of sciences,
  - as conics, linguistics, optics,
- or matters of practice,
  - as economics, politics, tactics
- linguistically, the meaning extends easily
  - to encompass both
    - the science of information
    - the practice of information processing.

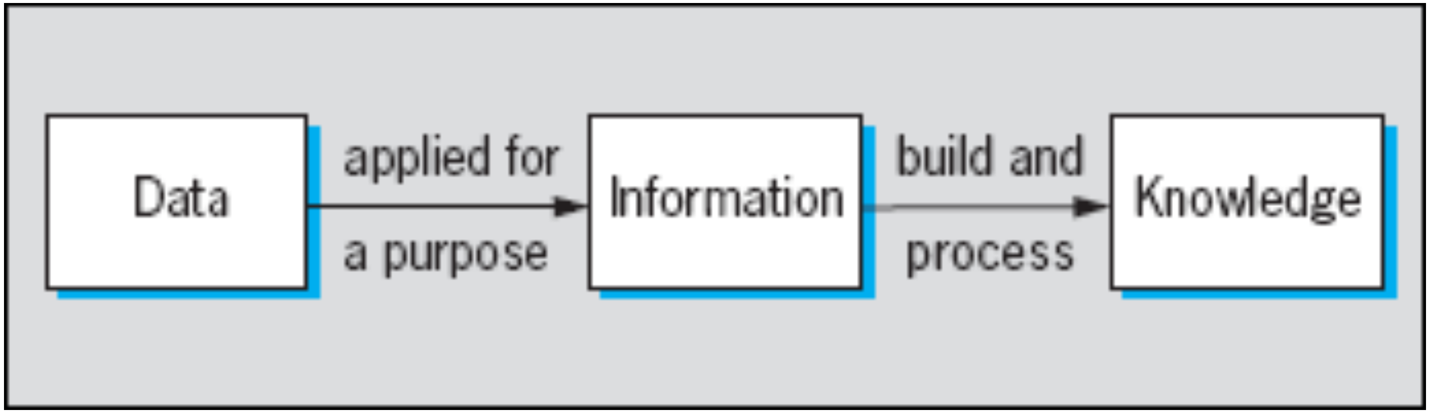
# Data - Information - Knowledge

- Data
  - unprocessed facts and figures without any added interpretation or analysis.
    - {The price of crude oil is $80 per barrel.}
- Information
  - data that has been interpreted so that it has meaning for the user.
    - {The price of crude oil has risen from $70 to $80 per barrel}
      - [gives meaning to the data and so is said to be information to someone who tracks oil prices.]

# Data - Information - Knowledge

- Knowledge
  - a combination of information, experience and insight that may benefit the individual or the organisation.
    - {When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per litre.}
      - [This is knowledge]
      - [insight: the capacity to gain an accurate and deep understanding of someone or something; an accurate and deep understanding]

# Converting data into information




- Data becomes information when it is applied to some purpose and adds value for the recipient.
  - For example a set of raw sales figures is data.
    - For the Sales Manager tasked with solving a problem of poor sales in one region, or deciding the future focus of a sales drive, the raw data needs to be processed into a sales report.
  - It is the sales report that provides information.

# Converting data into information

- Collecting data is expensive
  - you need to be very clear about why you need it and how you plan to use it.
  - One of the main reasons that organizations collect data is to monitor and improve performance.
    - if you are to have the information you need for control and performance improvement, you need to:
      - collect data on the indicators that really do affect performance
      - collect data reliably and regularly
      - be able to convert data into the information you need.

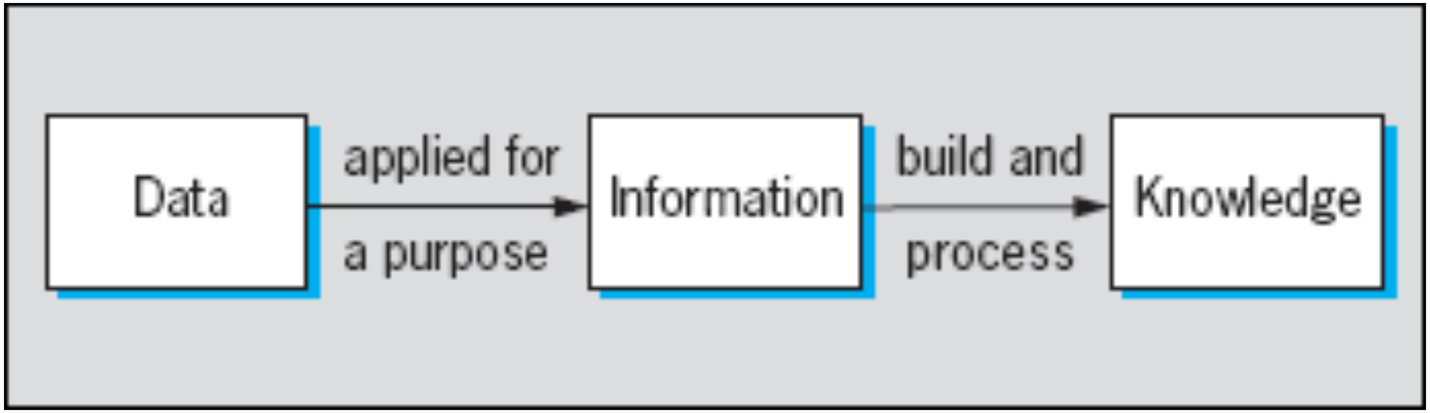
# Converting data into information

- To be useful, data must satisfy a number of conditions. It must be:
  - relevant to the specific purpose
  - complete
  - accurate
  - timely
    - data that arrives after you have made your decision is of no value

# Converting data into information

  - in the right format
    - information can only be analysed using a spreadsheet if all the data can be entered into the computer system
  - available at a suitable price
    - the benefits of the data must merit the cost of collecting or buying it.
- The same criteria apply to information.
  - It is important
    - to get the right information
    - to get the information right

# Converting information to knowledge




- Ultimately the tremendous amount of information that is generated is only useful if it can be applied to create knowledge within the organization.
- There is considerable blurring and confusion between the terms information and knowledge.

# Converting information to knowledge

- think of knowledge as being of two types:
  - Formal, explicit or generally available knowledge.
    - This is knowledge that has been captured and used to develop policies and operating procedures for example.
  - Instinctive, subconscious, tacit or hidden knowledge.
    - Within the organisation there are certain people who hold specific knowledge or have the 'know how'
      - {"I did something very similar to that last year and this happened….."}

# Converting information to knowledge

- Clearly, both types of knowledge are essential for the organisation.
- Information on its own will not create a knowledge-based organisation
  - but it is a key building block.
- The right information fuels the development of intellectual capital
  - which in turns drives innovation and performance improvement.

*** ÖZETLE:**

- Data (Veri): Henüz işlenmemiş ham bilgilerdir.
- Information (Maʼ lumat[1]): İşlenmiş, özetlenmiş, organize edilmiş, anlam kazanmış veridir.
- Knowledge (Bilgi, vukuf): Bir karara varmak için, maʼ lumatın deney, tecrübe, yorum, analiz ve sentezle zenginleştirilmiş hali

1. Maʼ lumat Arapça kökenli bir kelime olup çoğul nitelik taşır.
"Bu hakikatler artık çocukların bildikleri en basit malumat sırasına geçmiştir." - H. R. Gürpınar

# Analysis and Synthesis

The terms analysis and synthesis come from Greek

- they mean respectively "to take apart" and "to put together".
- These terms are in scientific disciplines from mathematics and logic to economy and psychology to denote similar investigative procedures.

Analysis is defined as the procedure by which we break down an intellectual or substantial whole into parts.

Synthesis is defined as the procedure by which we combine separate elements or components in order to form a coherent whole.

# DEFINITION(S) OF SYSTEM

A system can be broadly defined as an integrated set of elements that accomplish a defined objective.

People from different engineering disciplines have different perspectives of what a "system" is.

For example,

- software engineers often refer to an integrated set of computer programs as a "system"
- electrical engineers might refer to complex integrated circuits or an integrated set of electrical units as a "system"

As can be seen, "system" depends on one's perspective, and the "integrated set of elements that accomplish a defined objective" is an appropriate definition.

# Definition(s) of system

- A system is an assembly of parts where:
  - The parts or components are connected together in an organized way.
  - The parts or components are affected by being in the system (and are changed by leaving it).
  - The assembly does something.
  - The assembly has been identified by a person as being of special interest.
- Any arrangement which involves the handling, processing or manipulation of resources of whatever type can be represented as a system.
- Some definitions on online dictionaries
  - http://en.wikipedia.org/wiki/System
  - http://dictionary.reference.com/browse/systems
  - http://www.businessdictionary.com/definition/system.html

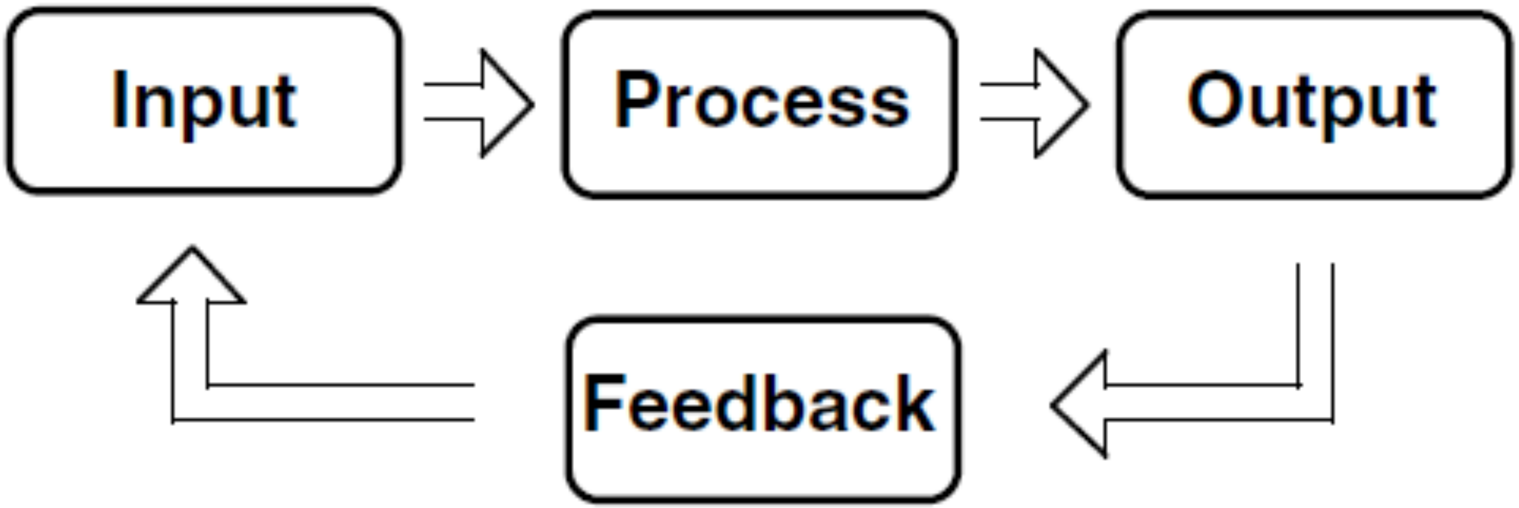
# Definition(s) of system

- A system is defined as multiple parts working together for a common purpose or goal.
- Systems can be large and complex
  - such as the air traffic control system or our global telecommunication network.
- Small devices can also be considered as systems
  - such as a pocket calculator, alarm clock, or 10-speed bicycle.

# Definition(s) of system

- Systems have inputs, processes, and outputs.
- When feedback (direct or indirect) is involved, that component is also important to the operation of the system.
- To explain all this, systems are usually explained using a model.
- A model helps to illustrate the major elements and their relationship, as illustrated in the next slide

# A systems model



# Information Systems

- The ways that organizations
  - Store
  - Move
  - Organize
  - Process

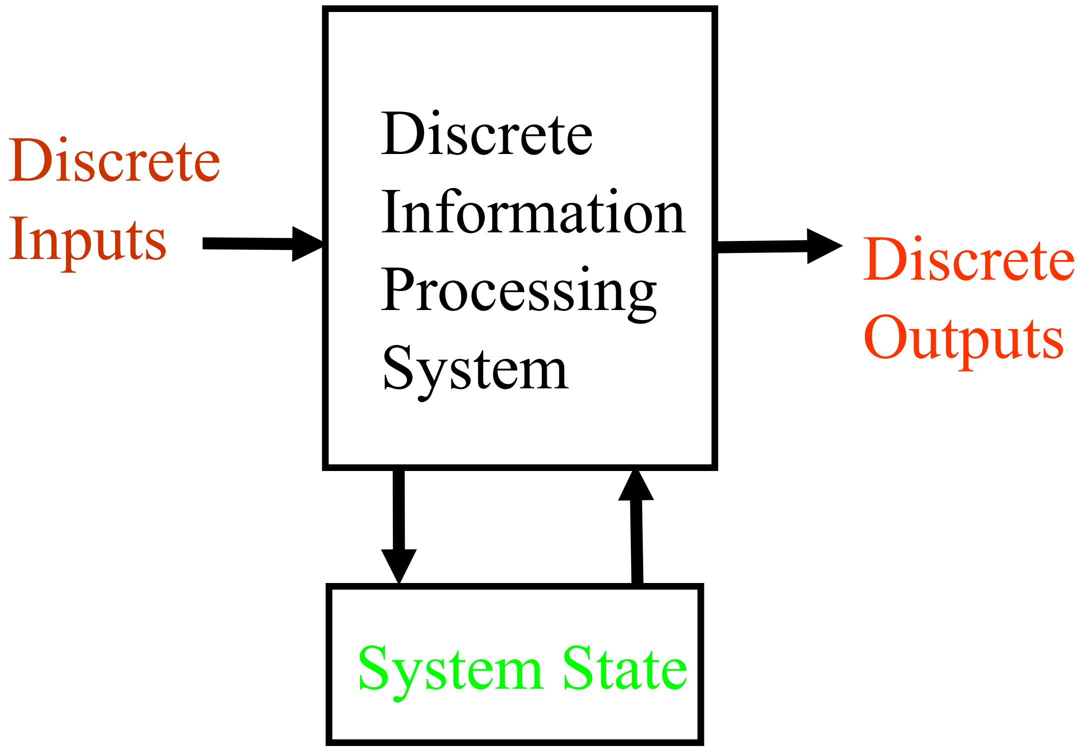
  their information

# Information Technology

- Components that implement information systems,
  - Hardware
    - physical tools: computer and network hardware, but also low-tech things like pens and paper
  - Software
    - (changeable) instructions for the hardware
  - People
  - Procedures
    - instructions for the people
  - Data/databases

# Digital System

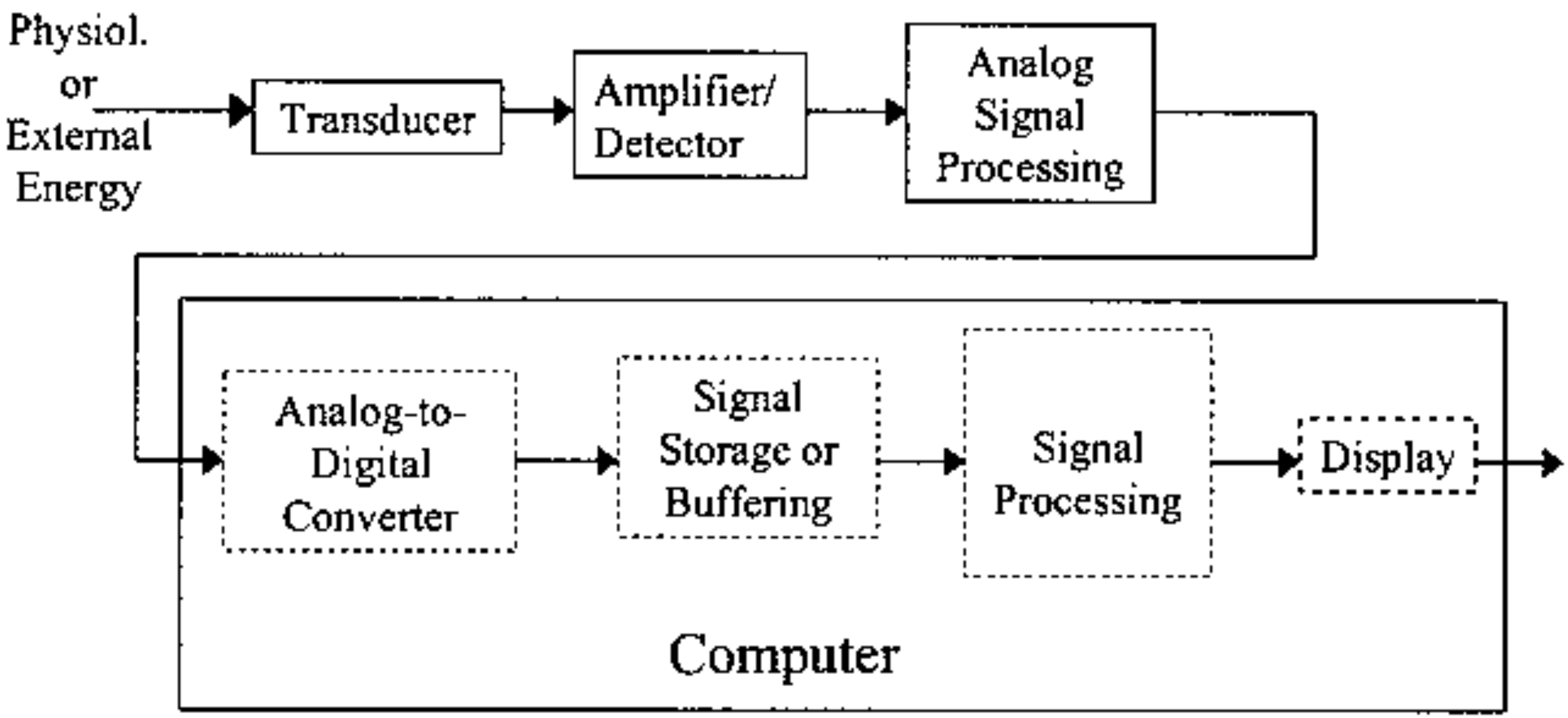
- **Takes a set of discrete information (inputs) and discrete internal information (system state) and generates a set of discrete information (outputs).**



# Signal

- An information variable represented by physical quantity.
- For digital systems, the variable takes on discrete values.
- Two level, or binary values are the most prevalent values in digital systems.
- Binary values are represented abstractly by:
  - digits 0 and 1
  - words (symbols) False (F) and True (T)
  - words (symbols) Low (L) and High (H)
  - and words On and Off.
- Binary values are represented by values or ranges of values of physical quantities

# A typical measurement system



# Transducers

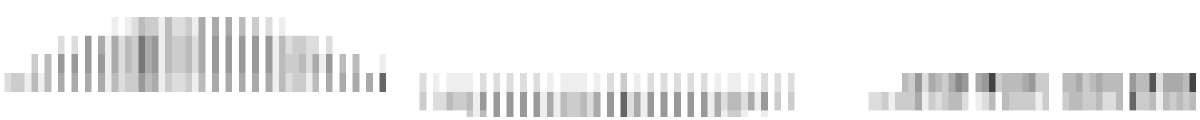
- A "transducer" is a device that converts energy from one form to another.
- In signal processing applications, the purpose of energy conversion is to transfer information, not to transform energy.
- In physiological measurement systems, transducers may be
  - input transducers (or sensors)
    - they convert a non-electrical energy into an electrical signal.
    - for example, a microphone.
  - output transducers (or actuators)
    - they convert an electrical signal into a non-electrical energy.
    - For example, a speaker.

- The analogue signal
  - a continuous variable defined with infinite precision

  is converted to a discrete sequence of measured values which are represented digitally
- Information is lost in converting from analogue to digital, due to:
  - inaccuracies in the measurement
  - uncertainty in timing
  - limits on the duration of the measurement
- These effects are called quantisation errors

- The continuous analogue signal has to be held before it can be sampled

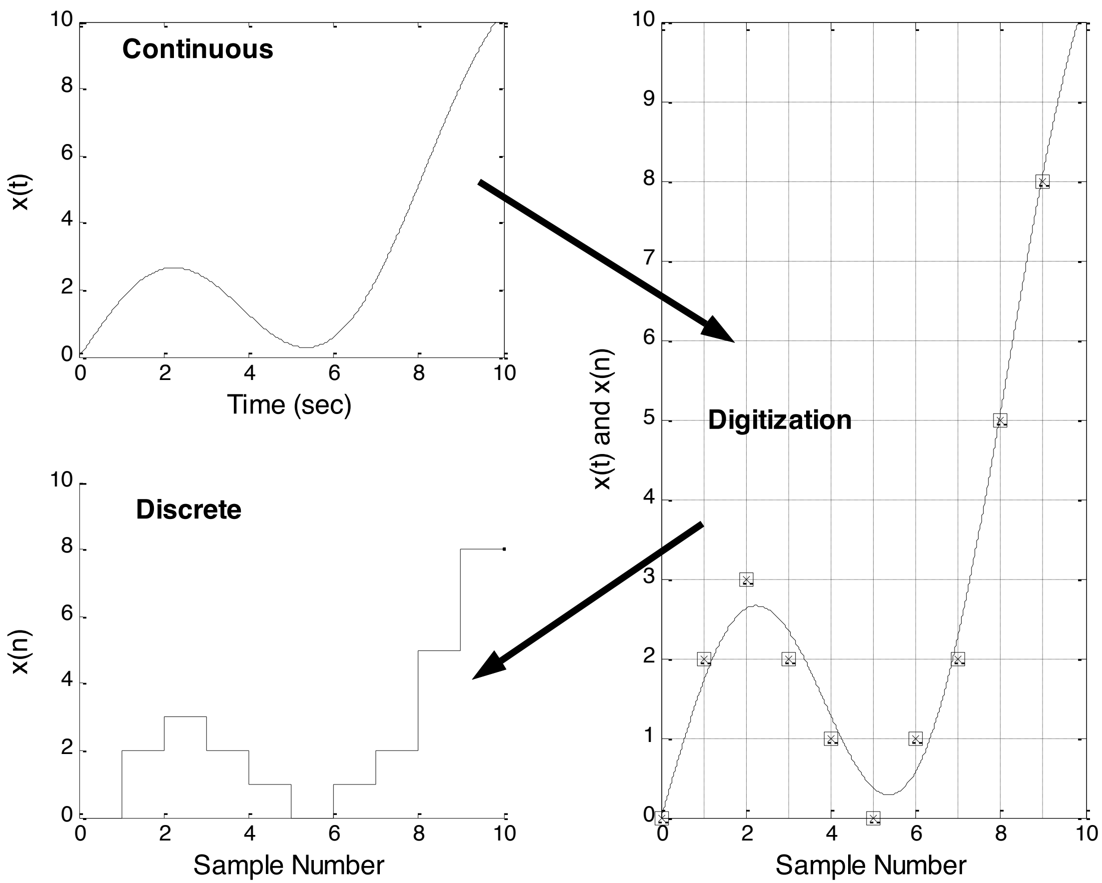

- Otherwise, the signal would be changing during the measurement
- Only after it has been held can the signal be measured, and the measurement converted to a digital value



# Signal Encoding: Analog-to Digital Conversion

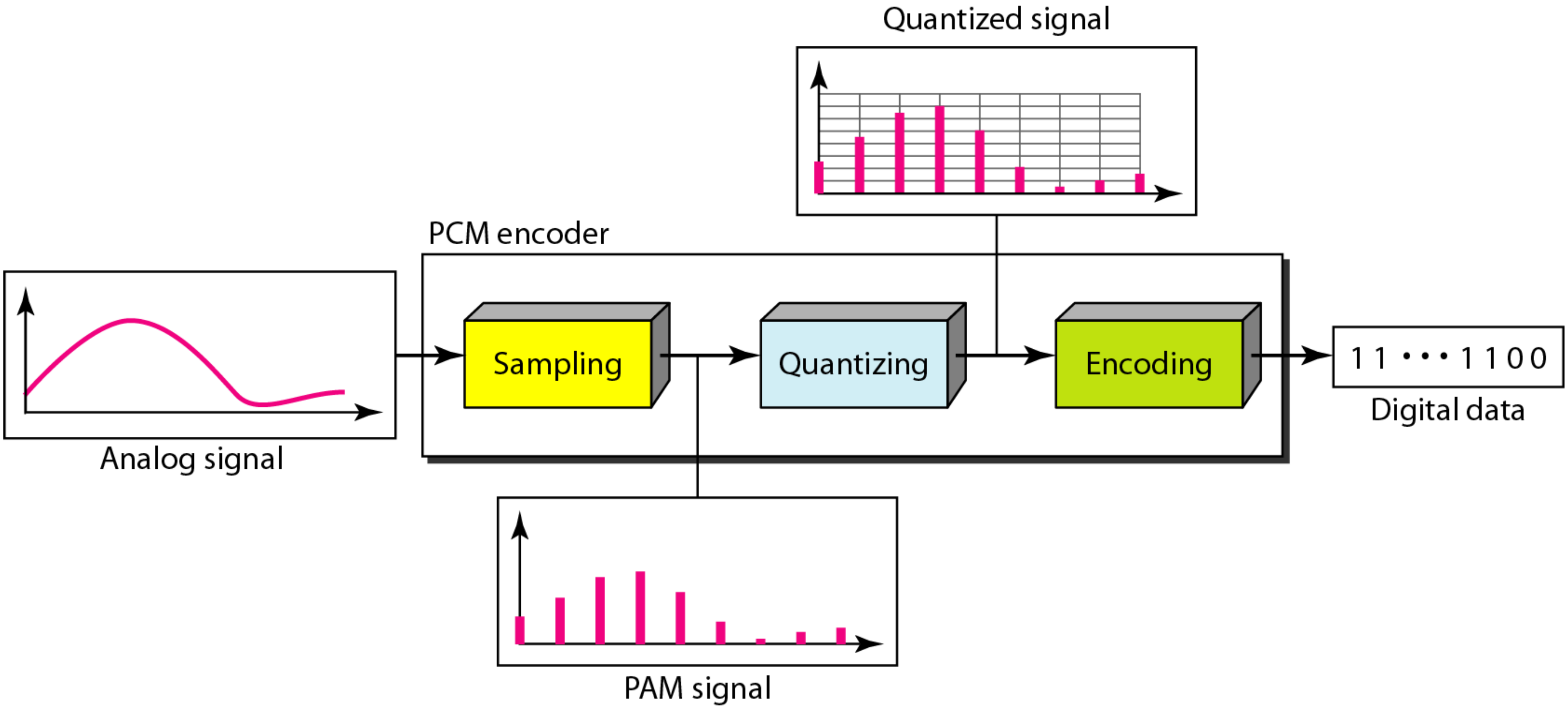
Continuous (analog) signal ↔ Discrete signal

$x$(t) = $f$(t) ↔ Analog to digital conversion ↔ $x$[n] = $x$ [1], $x$ [2], $x$ [3], ... $x$[n]



# Analog-to-Digital Conversion

- ADC consists of four steps to digitize an analog signal:
  1. Filtering
  2. Sampling
  3. Quantization
  4. Binary encoding
- Before we sample, we have to filter the signal to limit the maximum frequency of the signal as it affects the sampling rate.
- Filtering should ensure that we do not distort the signal, ie remove high frequency components that affect the signal shape.

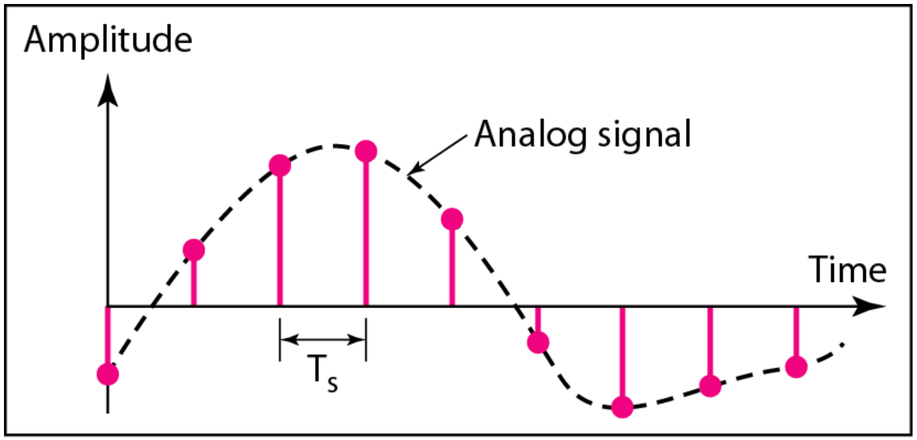

Quantized signal
PCM encoder
Sampling
Quantizing
Encoding
1 1 • • • 1 1 0 0
Digital data
Analog signal
PAM signal

# Sampling

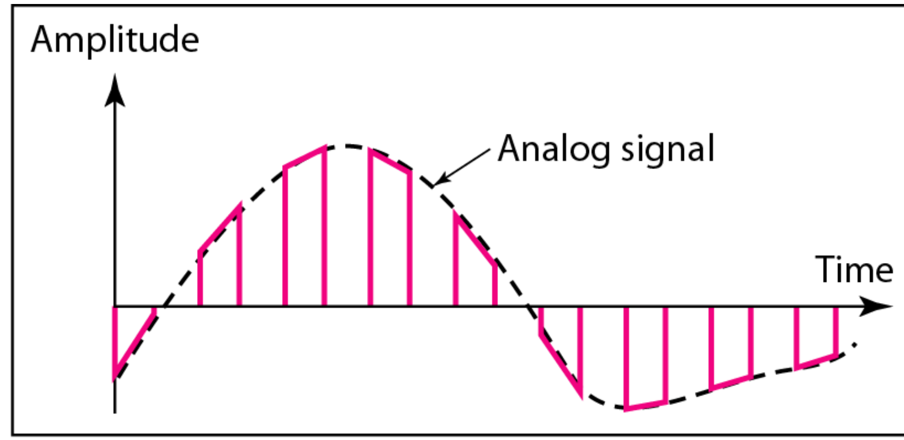
- The sampling results in a discrete set of digital numbers that represent measurements of the signal
  - usually taken at equal intervals of time
- Sampling takes place after the hold
  - The hold circuit must be fast enough that the signal is not changing during the time the circuit is acquiring the signal value
- We don't know what we don't measure
- In the process of measuring the signal, some information is lost

# Sampling

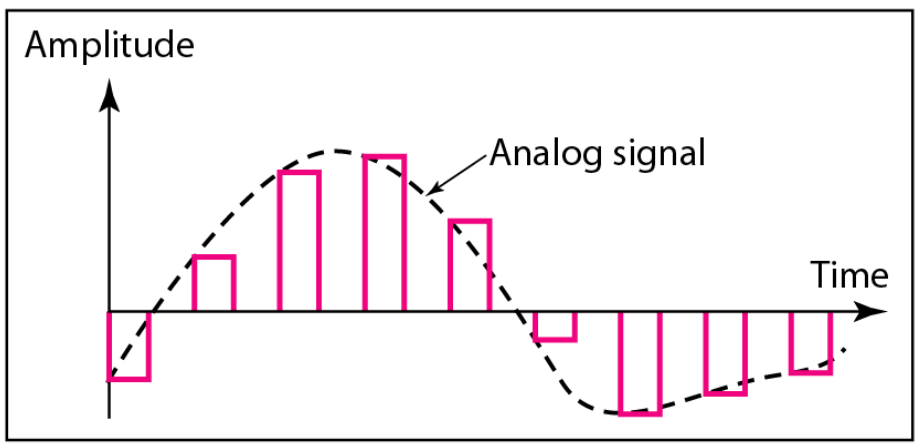
- Analog signal is sampled every $T_S$ secs.
- $T_s$ is referred to as the sampling interval.
- $f_s = 1/T_s$ is called the sampling rate or sampling frequency.
- There are 3 sampling methods:
  - Ideal - an impulse at each sampling instant
  - Natural - a pulse of short width with varying amplitude
  - Flattop - sample and hold, like natural but with single amplitude value
- The process is referred to as pulse amplitude modulation (PAM) and the outcome is a signal with analog (non integer) values



a. Ideal sampling

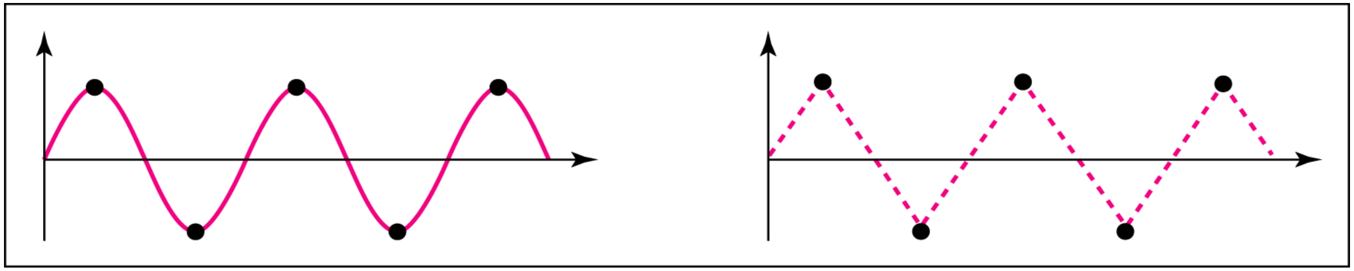


b. Natural sampling

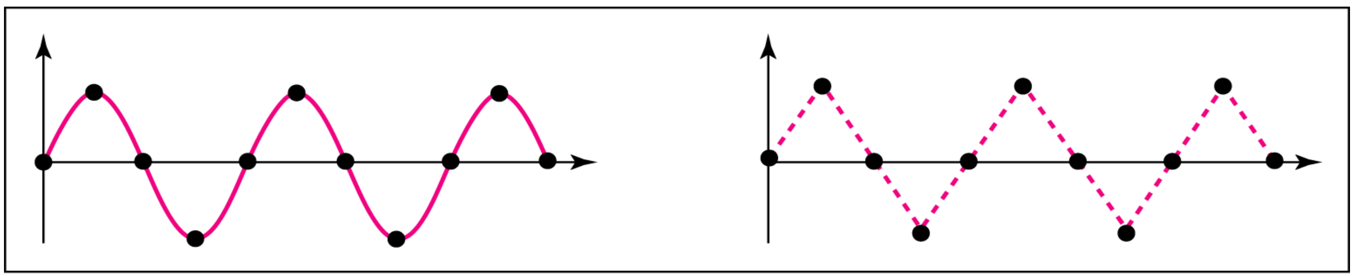


c. Flat-top sampling

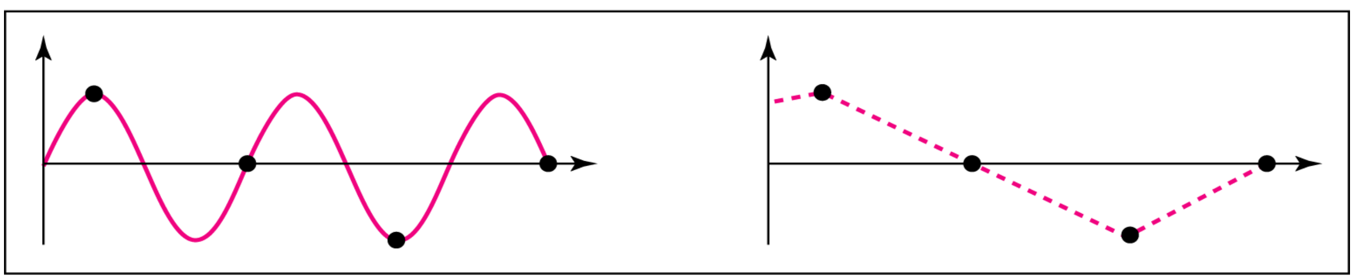
# Recovery of a sampled sine wave for different sampling rates



a. Nyquist rate sampling: $f_s = 2\,f$



b. Oversampling: $f_s = 4\,f$



c. Undersampling: $f_s = f$

# Sampling Theorem

$$F_s \geq 2f_m$$

**According to the Nyquist theorem, the sampling rate must be at least 2 times the highest frequency contained in the signal.**

# Quantization

- Sampling results in a series of pulses of varying amplitude values ranging between two limits: a min and a max.
- The amplitude values are infinite between the two limits.
- We need to map the *infinite* amplitude values onto a finite set of known values.
- This is achieved by dividing the distance between min and max into L zones, each of height Δ.

$$\Delta = (\max - \min)/L$$

# Quantization Levels

- The midpoint of each zone is assigned a value from 0 to L-1 (resulting in L values)
- Each sample falling in a zone is then approximated to the value of the midpoint.

# Quantization Zones

- Assume we have a voltage signal with amplitutes $V_{min}$=-20V and $V_{max}$=+20V.
- We want to use L=8 quantization levels.
- Zone width $\Delta = (20 - -20)/8 = 5$
- The 8 zones are: -20 to -15, -15 to -10, -10 to -5, -5 to 0, 0 to +5, +5 to +10, +10 to +15, +15 to +20
- The midpoints are: -17.5, -12.5, -7.5, -2.5, 2.5, 7.5, 12.5, 17.5

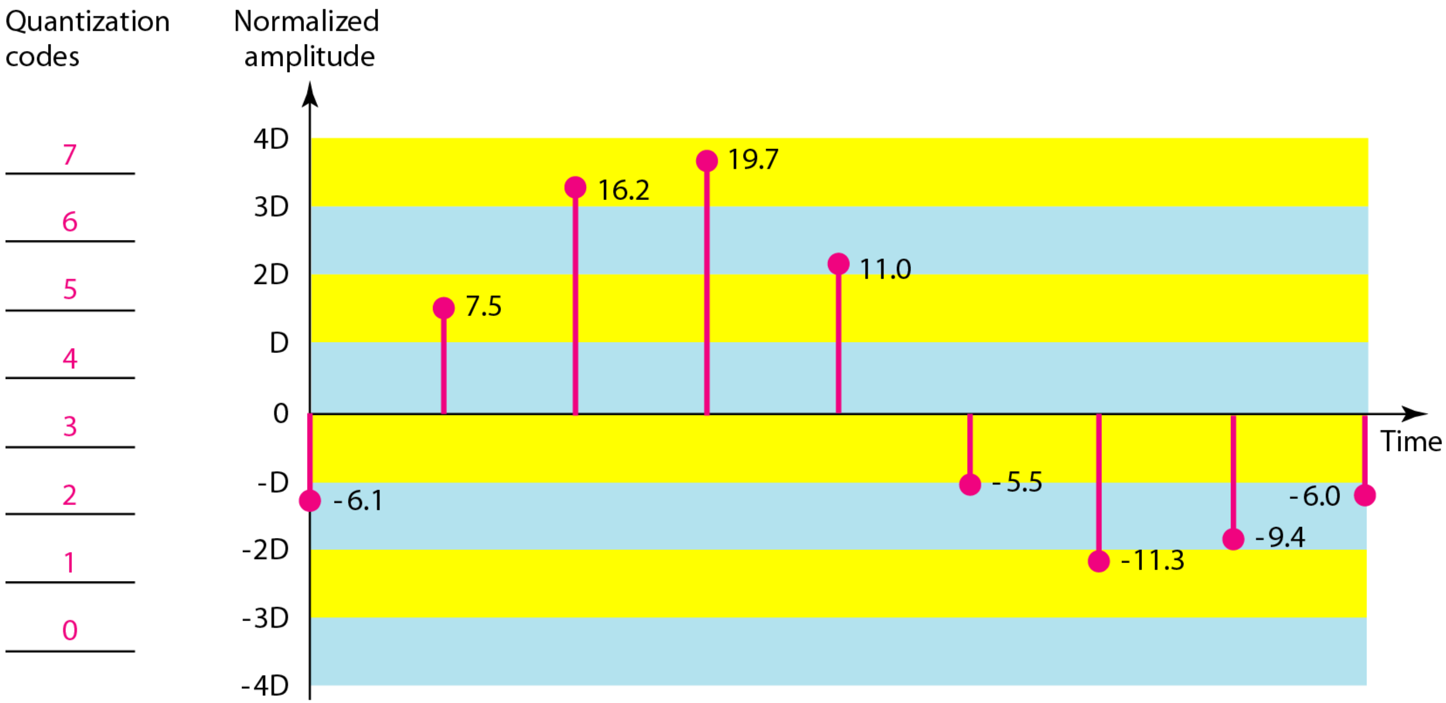
# Assigning Codes to Zones

- Each zone is then assigned a binary code.
- The number of bits required to encode the zones, or the number of bits per sample as it is commonly referred to, is obtained as follows:

$$n_b = \log_2 L$$

- Given our example, $n_b = 3$
- The 8 zone (or level) codes are therefore: 000, 001, 010, 011, 100, 101, 110, and 111
- Assigning codes to zones:
  - 000 will refer to zone -20 to -15
  - 001 to zone -15 to -10, etc.

# Quantization and encoding of a sampled signal




| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Normalized PAM values | -1.22 | 1.50 | 3.24 | 3.94 | 2.20 | -1.10 | -2.26 | -1.88 | -1.20 |
| Normalized quantized values | -1.50 | 1.50 | 3.50 | 3.50 | 2.50 | -1.50 | -2.50 | -1.50 | -1.50 |
| Normalized error | -0.38 | 0 | +0.26 | -0.44 | +0.30 | -0.40 | -0.24 | +0.38 | -0.30 |
| Quantization code | 2 | 5 | 7 | 7 | 6 | 2 | 1 | 2 | 2 |
| Encoded words | 010 | 101 | 111 | 111 | 110 | 010 | 001 | 010 | 010 |

# Quantization Error

- When a signal is quantized, we introduce an error
  - the coded signal is an approximation of the actual amplitude value.
- The difference between actual and coded value (midpoint) is referred to as the quantization error.
- The more zones, the smaller $\Delta$
  - which results in smaller errors.
- BUT, the more zones the more bits required to encode the samples
  - higher bit rate

# Analog-to-digital Conversion

**Example** An 12-bit analog-to-digital converter (ADC) advertises an accuracy of ± the least significant bit (LSB). If the input range of the ADC is 0 to 10 volts, what is the accuracy of the ADC in analog volts?

Solution:

If the input range is 10 volts then the analog voltage represented by the LSB would be:

$$V_{LSB} = \frac{V_{max}}{2^{\text{Nu bits}}} = \frac{10}{2^{12}} = \frac{10}{4096} = .0024 \text{ volts}$$

Hence the accuracy would be ± 0.0024 volts.

# Sampling related concepts

- Over/exact/under sampling
- Regular/irregular sampling
- Linear/Logarithmic sampling
- Aliasing
- Anti-aliasing filter
- Image
- Anti-image filter

# Steps for digitization/reconstruction of a signal

- Band limiting (LPF)
- Sampling / Holding
- Quantization
- Coding

*These are basic steps for A/D conversion*

- D/A converter
- Sampling / Holding
- Image rejection

*These are basic steps for reconstructing a sampled digital signal*

# Digital data: end product of A/D conversion and related concepts

- Bit: least digital information, binary 1 or 0
- Nibble: 4 bits
- Byte: 8 bits, 2 nibbles
- Word: 16 bits, 2 bytes, 4 nibbles
- Some jargon:
  - integer, signed integer, long integer, 2s complement, hexadecimal, octal, floating point, etc.


MP3 Audio
Recording
ADC
00000
00000
DSP
00000
MEMORY
DAC
00000
00000
Playback

# Measures of capacity and speed in Computers

Special Powers of 10 and 2 :

| | | | | |
|---|---|---|---|---|
| • Kilo- (K) | = 1 thousand | = $10^3$ | and | $2^{10}$ |
| • Mega- (M) | = 1 million | = $10^6$ | and | $2^{20}$ |
| • Giga- (G) | = 1 billion | = $10^9$ | and | $2^{30}$ |
| • Tera- (T) | = 1 trillion | = $10^{12}$ | and | $2^{40}$ |
| • Peta- (P) | = 1 quadrillion | = $10^{15}$ | and | $2^{50}$ |

**Whether a metric refers to a power of ten or a power of two typically depends upon what is being measured.**

# Example

- Hertz = clock cycles per second (frequency)
  - 1MHz = 1,000,000Hz
  - Processor speeds are measured in MHz or GHz.
- Byte = a unit of storage
  - 1KB = $2^{10}$ = 1024 Bytes
  - 1MB = $2^{20}$ = 1,048,576 Bytes
  - Main memory (RAM) is measured in MB
  - Disk storage is measured in GB for small systems, TB for large systems.

# Measures of time and space

- Milli- (m) = 1 thousandth = $10^{-3}$
- Micro- (μ) = 1 millionth = $10^{-6}$
- Nano- (n) = 1 billionth = $10^{-9}$
- Pico- (p) = 1 trillionth = $10^{-12}$
- Femto- (f) = 1 quadrillionth = $10^{-15}$