

CSH1F2 Introduction to Computer Science

Week 5: Data Manipulation and Visualization

Author: Lecturer Team



Data Manipulation and Visualization

1. Introduction to Data

Data

- ▶ Data is a collection of facts : numbers, words, measurements, observations or even just descriptions of things
- ▶ Data can be **qualitative** or **quantitative**
- ▶ Quantitative data can be **discrete** or **continuous**
- ▶ Discrete data : 5,35, 56, ...
- ▶ Continuous data: 3.99 ;

Introduction to Data [1]

- ▶ collecting and analyzing data
- ▶ many ways data impacts lives
- ▶ how it can be used.
- ▶ where and how data is collected,
- ▶ who is collecting it, and
- ▶ how they are using it.

Introduction to Data [2]

- ▶ In computing, we're interested in:
 - where data comes from,
 - what structure or formats it comes in, and
 - what kind of knowledge or information we can extract from that data using computational tools.
- ▶ Who is generating the data?
- ▶ Where is the data being stored or saved? Who owns it?

Introduction to Data [3]

- ▶ In Computer Science, sometimes we can have the computer itself generate data for us.
- ▶ There are other kinds of data that can't be generated by the computer.
- ▶ In particular, data about people and how they act in the real world is hard to capture without just asking them.
- ▶ So that's what a lot of tools online do. They try to capture people's responses to things because the data, in aggregate, might contain useful information that could be extracted.
- ▶ Example :
<http://www.zimbio.com/quiz/3FINH9tImJ4/How+Much+Left+Right+Brained+Person>.

Introduction to Data [4]


- ▶ That “dumb” online quiz is an example.
- ▶ These quizzes ask people to reveal things about themselves, their preferences, likes and dislikes. This is data!
- ▶ While these online quizzes are probably innocuous, some interesting things about people could probably be discovered if the data were analyzed.

Online quiz looks like

5 of 13

W 7 of 13

f
8 c



You Are
50% Left-Brained,
50% Right-Brained

Logic and analysis rule your mental processor. You are efficient, strategic, and always in control. When things don't go according to plan, it really throws you off, but you have the determination to persevere through life's challenges. And if perseverance doesn't cut it, you have the wit and wisdom to talk your way out of difficult situations. You are a rock.

ZIMBIO

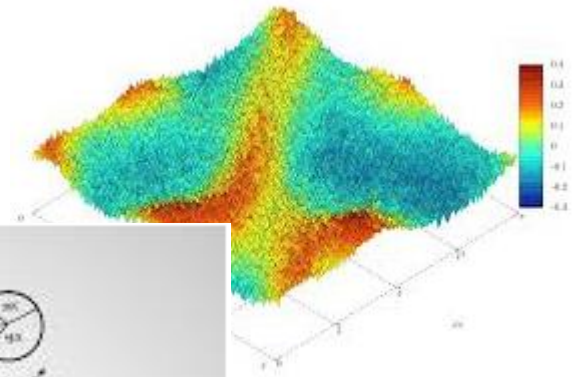
How to use data ?

[illegible]

Data Manipulation and Visualization

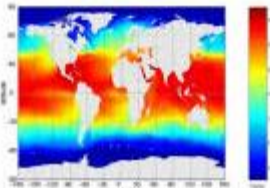
- 1. Introduction to Data**
- 2. Finding Trends with Visualizations**

Data Visualization

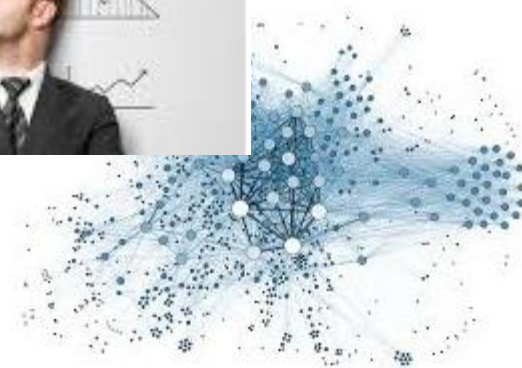


Gambar berikut ini menunjukkan suhu permukaan laut pada bulan Juli tahun 1987

Sepuluh dari ribuan titik data diringkas dal satu gambar



Kompor Data Mining



Basic Data Visualization Techniques

- Bar Graphs
- Histogram
- Line Graphs
- Pie Charts
- Scatter Plots
- Time series graphs



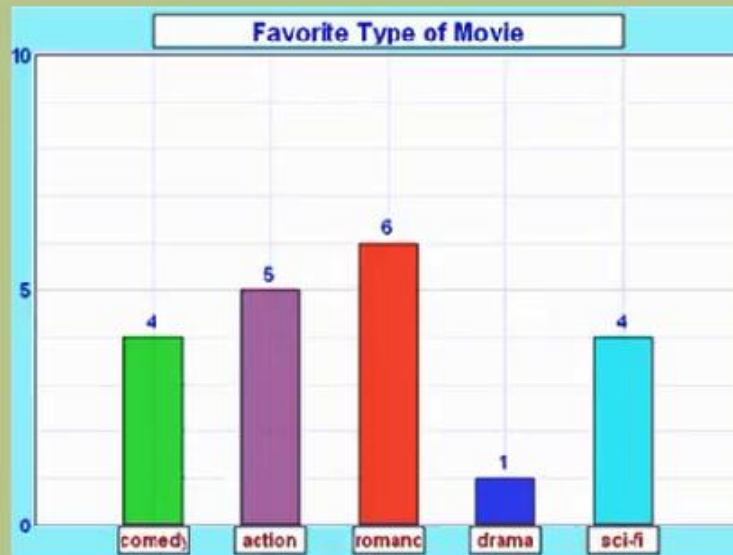
Bar Graphs

A **Bar Graph** (also called Bar Chart) is a graphical display of data using bars of different heights.

Also known as a Pareto Diagram, a bar graph can be horizontal or vertical. Each axis is labeled with either a categorical or a numerical variable. The bars' heights are scaled according to their values and the bars can be compared to each other.

Imagine you just did a survey of your friends to find which kind of movie they liked best:

Table: Favorite Type of Movie



Comedy	Action	Romance	Drama	Horror
4	5	6	1	4

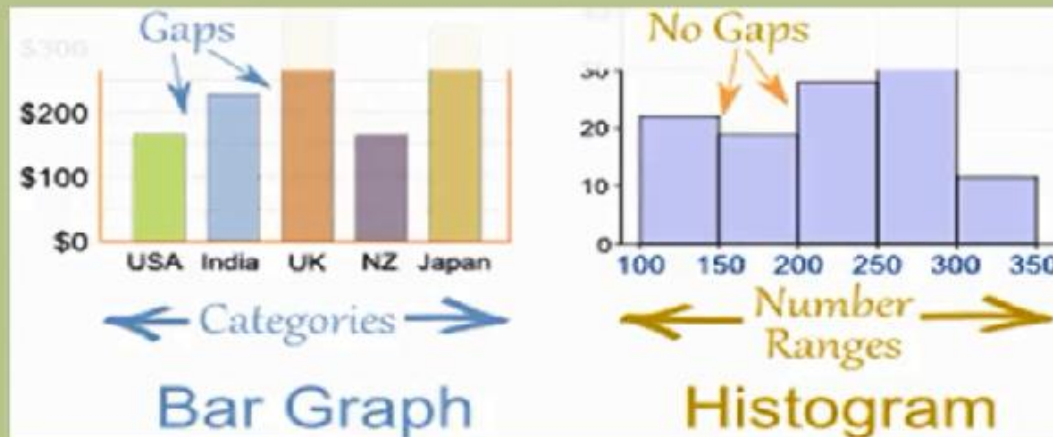
It is a really good way to show relative sizes: we can see which types of movie are most liked, and which are least liked, at a glance.

Histogram

Bar Graphs are good when your data is in **categories** (such as "Comedy", "Drama", etc).

But when you have continuous data (such as a person's height) then use a Histogram.

It is best to leave gaps between the bars of a Bar Graph, so it doesn't look like a Histogram.



Histogram

This is a kind of graph that also uses bars. Ranges of values are listed at the bottom and these are called 'classes.' Taller bars represent the classes with greater frequencies.

Line Graphs

Line Graph - A graph that shows information that is connected in some way (such as change over time)

You are learning facts about astrology, and each day you do a short test to see how good you are. These are the results:
You seem to be improving!



Table: Facts I got Correct

Day 1	Day 2	Day 3	Day 4
3	4	12	15

Comparing various sets of data can be complicated, but line graphs make it easy. The plotted peaks and dips on the grid allow you to monitor and compare improvement and decline.

Pie Charts

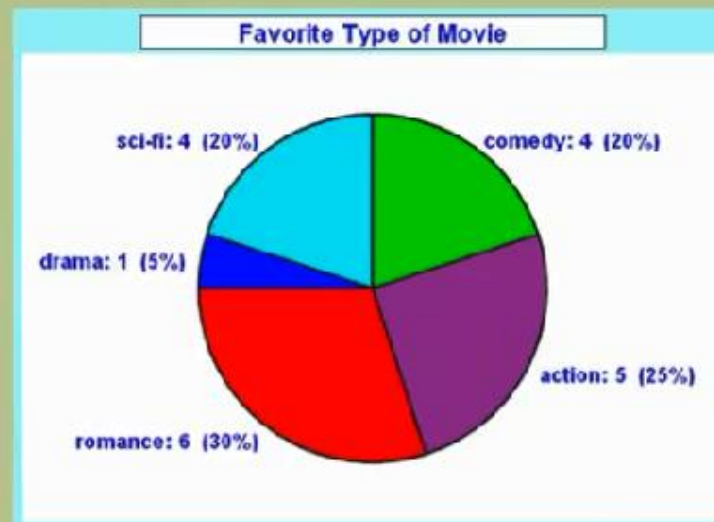
A special chart that uses "pie slices" to show relative sizes of data.

Imagine you just did a survey of your friends to find which kind of movie they liked best.

You could show that by this pie chart:

Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	Horror
4	5	6	1	4



Pie Charts

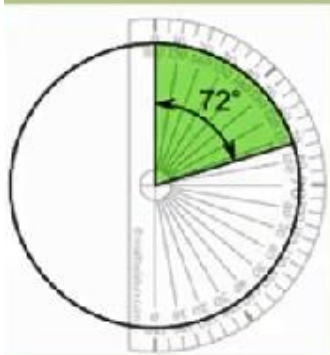
First, put your data into a table (like above), then add up all the values to get a total:

Next, divide each value by the total and multiply by 100 to get a percent:

Comedy	Action	Romance	Drama	Horror	TOTAL
4	5	6	1	4	20
$4/20 = 20\%$	$5/20 = 25\%$	$6/20 = 30\%$	$1/20 = 5\%$	$4/20 = 20\%$	100%

Now you need to figure out how many degrees for each "pie slice".

A Full Circle has 360 degrees, so we do this calculation:



Comedy	Action	Romance	Drama	Horror	TOTAL
4	5	6	1	4	20
$4/20 = 20\%$	$5/20 = 25\%$	$6/20 = 30\%$	$1/20 = 5\%$	$4/20 = 20\%$	100%
$4/20 \times 360^\circ = 72^\circ$	$5/20 \times 360^\circ = 90^\circ$	$6/20 \times 360^\circ = 108^\circ$	$1/20 \times 360^\circ = 18^\circ$	$4/20 \times 360^\circ = 72^\circ$	360°

Then use your protractor to measure the degrees of each sector.

Here I show the first sector ... you can do the rest!

Scatter Plots

A Scatter (XY) Plot has points that show the relationship between two sets of data. It displays paired data using the vertical y axis and the horizontal x axis.
(The data is plotted on the graph as "Cartesian (x,y) Coordinates")

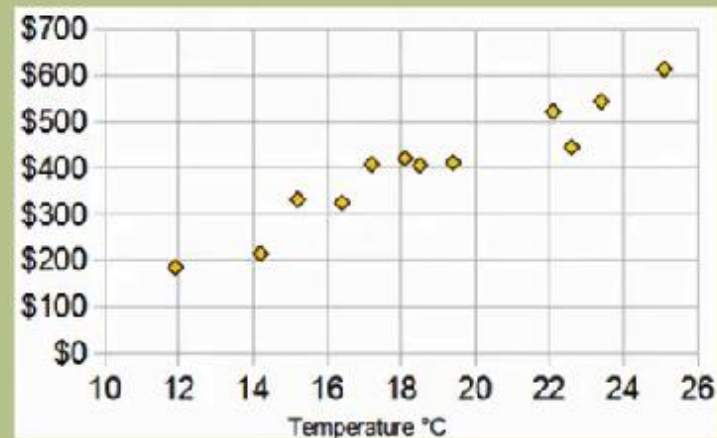
Example:

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

It is now easy to see that **warmer weather leads to more sales.**

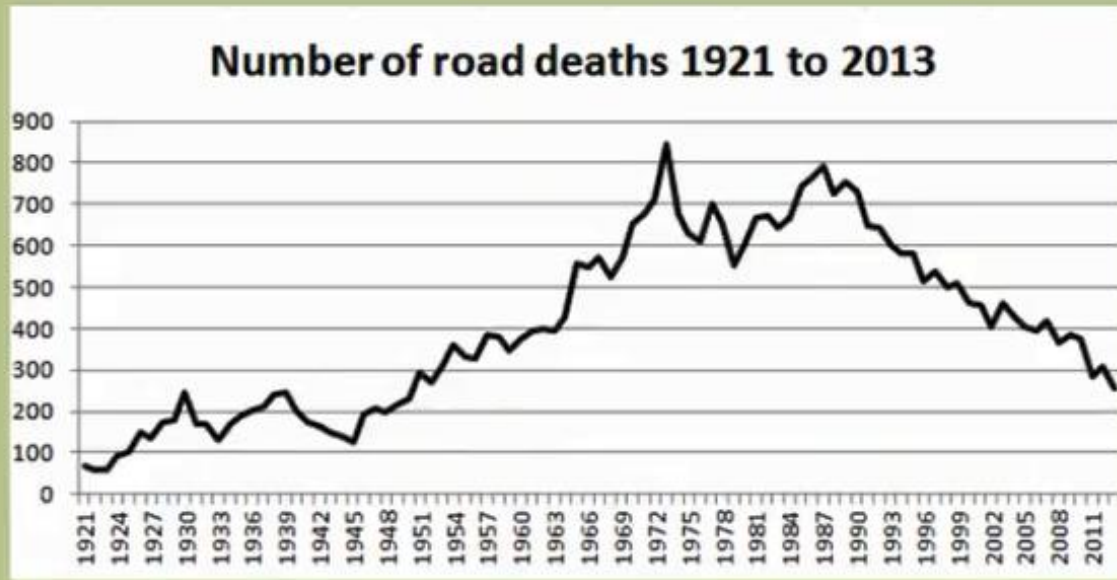
Ice Cream Sales vs Temperature

Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408



Time series graphs

Data is displayed in a time series graph at various time-points. The vertical axis is for data values while the horizontal axis shows time. This kind of graph can be used for showing trends passing through a time period.



Finding Trends with Visualizations

- ▶ We can use the Google Trends tool in order to visualize historical search data:
- ▶ identify interesting trends or
- ▶ patterns

You can try these links :

<https://www.google.com/trends/>

<https://support.google.com/trends#topic=4365530>

Finding Trends with Visualizations

- ▶ why we separate the **what** from the **why** when looking at data.
- ▶ The main purpose here is to **raise awareness of the assumptions** that we (all people) make when looking at data and try to call them out.
- ▶ Analyzing and interpreting data will typically require some assumptions to be made about the accuracy of the data and the cause of the relationships observed within it.

Finding Trends with Visualizations

- ▶ When decisions are made based on a collection of data, they will often rest just as much on that set of assumptions about the data as the data itself.
- ▶ Identifying and validating (or disproving) assumptions is therefore an important part of data analysis.

Activity 1 – will be for Session 6

- ▶ Create data visualization from the data below:

No.	NIM	Quiz	Tugas	UTS	UAS	Nilai Akhir
1	613091003	80	80	56	78	69.6
2	613120057	30	56	64	45	52.2
3	613090071	69	78	67	35	55.5
4	613120058	87	57	78	67	72.4
5	613120067	34	67	89	87	80.5
6	613120059	34	35	35	45	38.9
7	613120050	67	64	46	35	45.5
8	613120100	35	35	46	46	43.8
9	613120048	57	25	46	56	49
10	613120044	35	89	46	67	57.6

Activity 1 (cont...)

- ▶ You can create some data visualization for these following information:
 - The number of students which obtain quiz score below a particular value
 - The number of students which obtain quiz score >70
 - The number of students for each particular range
 - Etc.
- Data visualisation can be an appropriate graph
- See the video: Data & Medicine

Data Manipulation and Visualization

- 1. Introduction to Data**
- 2. Finding Trends with Visualizations**
- 3. Assumptions**

Assumptions [1]

- ▶ consider carefully the assumptions we make when interpreting data and data visualizations.
- ▶ how the Google Flu Trends project tried and failed to use search trends to predict flu outbreaks.
- ▶ See the video : Google Trend

Assumptions [2]

The most important points about Google Flu trends can be found below:

- Google Flu Trends worked well in some instances but often over-estimated, under-estimated, or entirely missed flu outbreaks. A notable example occurred when Google Flu Trends largely missed the outbreak of the H1N1 flu virus.
- Just because someone is reading about the flu doesn't mean they actually have it.
- Some search terms like "high school basketball" might be good predictors of the flu one year but clearly shouldn't be used to measure whether someone has the flu.

Assumptions [3]

- ▶ In general, many terms may have been good predictors of the flu for a while only because, like high school basketball, they are more searched in the winter when more people get the flu.
- ▶ Google began recommending searches to users, which skewed what terms people searched for. As a result, the tool was measuring Google-generated suggested searches as well, which skewed results.

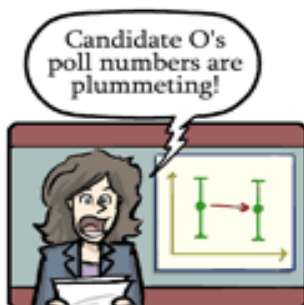
Assumptions [4]

- ▶ Digital Divide: access to technology differs widely by personal characteristics like race and income.
- ▶ A widespread assumption that data collected online is representative of the population at large.

Dear News Media,

When reporting poll results, please keep in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.



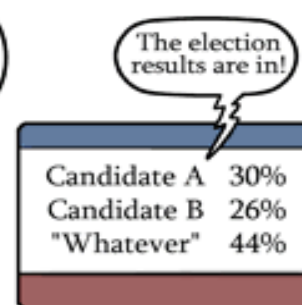
2. Scientific facts are not determined by public opinion polls.



3. A poll taken of your viewers/internet users is not a scientific poll.



4. What if all polls included the option "Don't care"?



Signed,

-Someone who took a
 basic statistics course.

JORGE CHAM © 2010

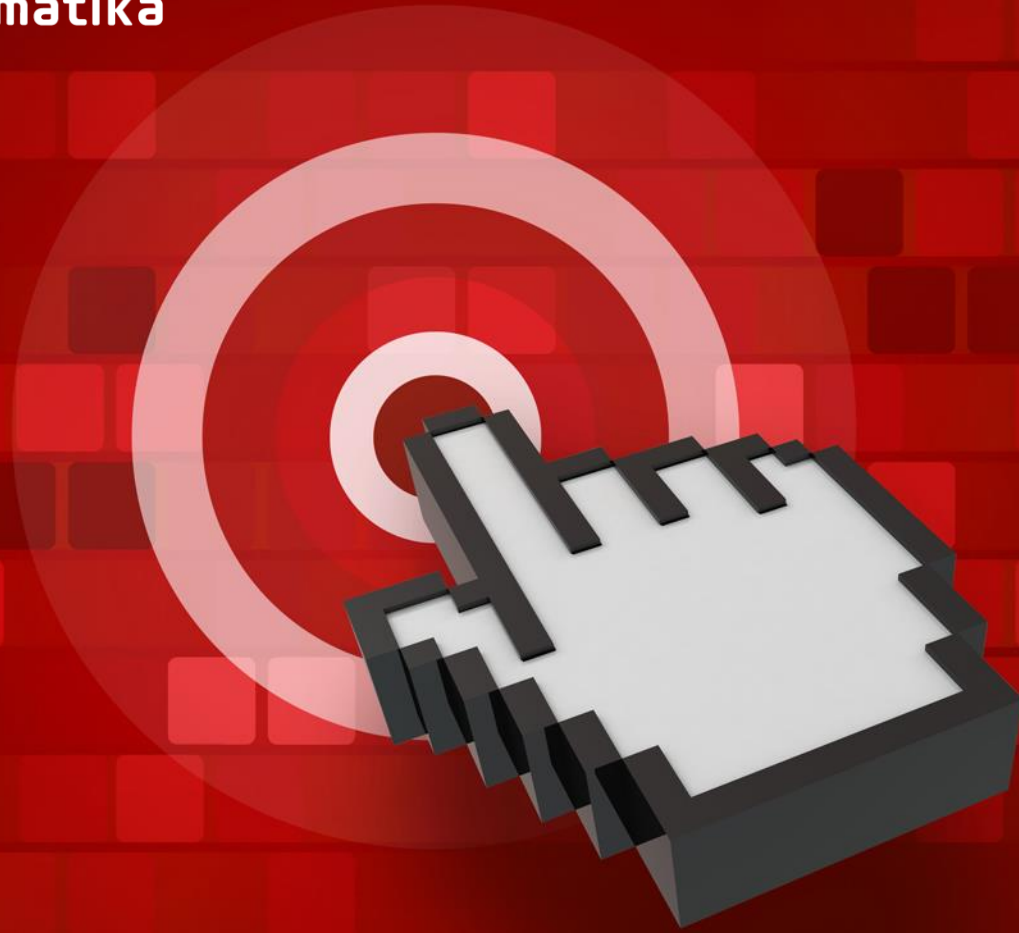
WWW.PHDCOMICS.COM

Worksheet

1. Describe various data with their examples
2. Describe some data visualization technique
(minimum 3 techniques)



Fakultas Informatika
School of Computing
Telkom University



THANK YOU