

# Data dan Variabel

---

Sukmawati Nur Endah – Universitas Diponegoro

# Pengertian Data

## ■ **Data (dataset):**

- Kumpulan dari data objek yang merepresentasikan sebuah entitas (atribut)

## ■ **Data objek** disebut juga sebagai record, point, sample dan instance

## ■ **Atribut** merepresentasikan karakteristik sebuah data objek

- Misalnya : untuk data mahasiswa, terdapat attribute Nama, NIM, Usia, Jenis Kelamin, dll
- Atribut disebut juga variabel atau fitur



# Contoh Dataset dan Bagiannya

## Dataset Bunga Iris

- Deskripsi dataset bunga Iris

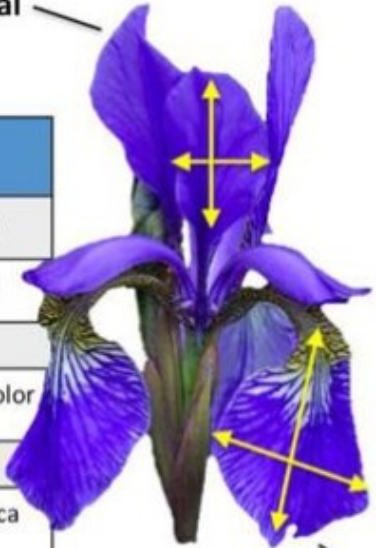
- Atribut : Sepal length, Sepal width, Petal length, dan Petal width
- Jumlah data : 150
- Kelas : Iris-setosa (50), Iris-versicolor (50), Iris-virginica (50)

**Samples**  
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

**Features**  
(attributes, measurements, dimensions)

**Class labels**  
(targets)



The diagram shows a purple Iris flower. A label 'Petal' points to the upper petals, and a label 'Sepal' points to the lower sepals. Yellow arrows indicate measurements: a vertical arrow on a petal for length, a horizontal arrow on a petal for width, and a diagonal arrow on a sepal for length.

# Tipe Atribut

## ■ Nominal

- Berupa kategori, misal jenis kelamin, status perkawinan, agama, dll

## ■ Binary

- Atribut nominal dengan hanya 2 nilai yaitu 0 dan 1

## ■ Ordinal

- Nilai yang merepresentasikan urutan, misal nilai mata kuliah

## ■ Numeric

- Quantity (bilangan integer atau real)
- Interval (ukuran skala unit, misal suhu, tanggal)
- Ratio (Misal panjang, harga, umur)

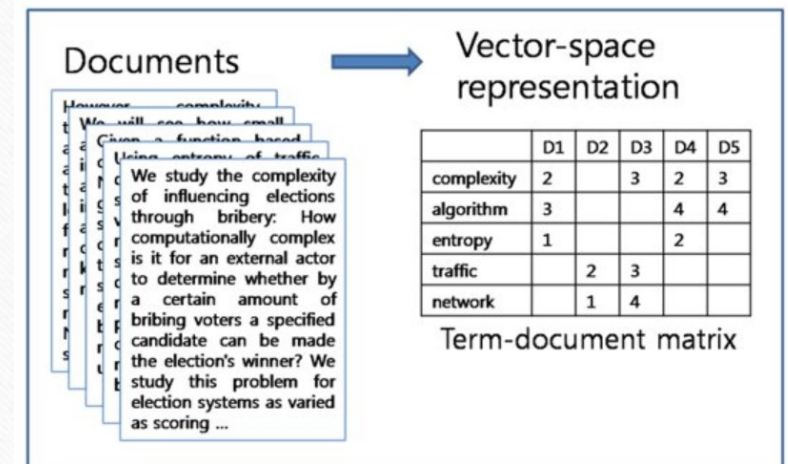
# Jenis-jenis Data

## Bentuk Record

- Data matrix
- Document data (document-term matrix)
- Transaction data

ID	Item
B1	Buku
B2	Pensil
B3	Spidol
B4	Pena

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes

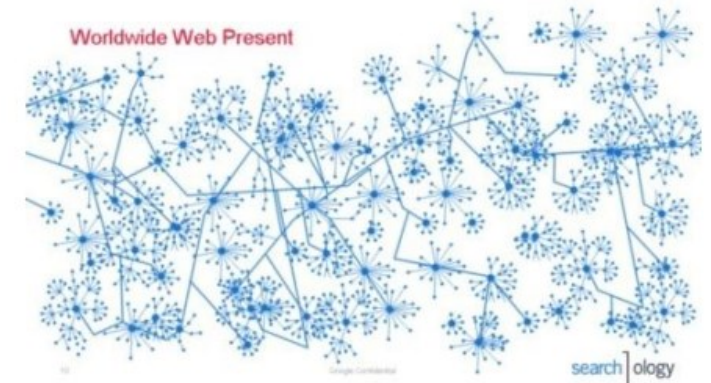




# Jenis-jenis Data

## ■ Bentuk Graph dan Network

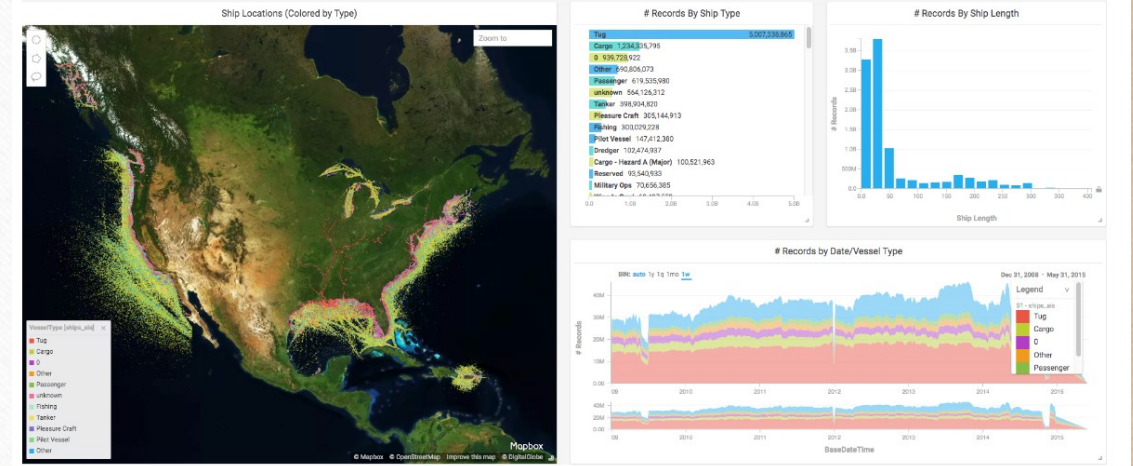
- ▷ World Wide Web
- ▷ Molecular Structures
- ▷ Social Networks



# Jenis-jenis Data

## Ordered

- Video data
- Spatio-Temporal data
- Sequential data
- Genetic sequence data



Sumber gambar : <https://www.omnisci.com/technical-glossary/spatial-temporal>

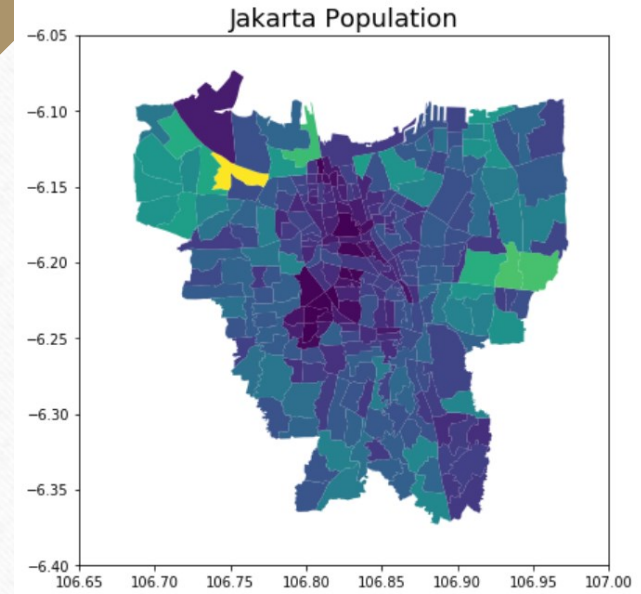
```
GGTTCCGCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



# Jenis-jenis Data

## ■ Spatial, image dan multimedia

- Spatial data (maps)
- Image data
- Voice data
- Video data



Sumber gambar : <https://towardsdatascience.com/spatial-data-solution-for-city-planning-in-indonesia-understanding-the-geodataframe-f50d58e6c9f2>





# Jenis data

---

- **Menurut cara memperolehnya**

- **Data Primer**

- data yang diambil langsung dari obyek penelitian atau merupakan data yang berasal dari sumber asli atau pertama
    - Pencarian data primer dilakukan dengan cara wawancara atau interview langsung dengan responden, melalui telepon, email instrumennya dalam bentuk kuisisioner

- **Data Sekunder**

- data yang tidak didapatkan secara langsung dari objek penelitian, melainkan data yang berasal dari sumber yang telah dikumpulkan oleh pihak lain.

# Jenis data

---

- Lanjutan Menurut Cara Memperolehnya
  - **Data Tersier**
    - Data yang berasal dari sumber-sumber yang diragukan keabsahannya seperti Wikipedia, Ensiklopedia cetak dan lain-lain
    - Sebaiknya tidak digunakan dalam penelitian



# Jenis data

---

- **Berdasarkan sumber data**

- **Data Internal**

- data yang menggambarkan situasi dan kondisi pada suatu organisasi secara internal.
    - Misal : data keuangan, data pegawai, data produksi, dan sebagainya.

- **Data Eksternal**

- data yang menggambarkan situasi serta kondisi yang ada di luar organisasi.
    - Contohnya adalah data jumlah penggunaan suatu produk pada konsumen, tingkat preferensi pelanggan, persebaran penduduk, dll

# Jenis Data

---

- **Berdasarkan bentuk data**

- **Data Kuantitatif**

- data yang dipaparkan dalam bentuk angka-angka.
    - Misalnya adalah jumlah pembeli pada saat menjelang hari raya Idul Fitri, tinggi badan

- **Data Kualitatif**

- data yang disajikan dalam bentuk kata-kata yang mengandung makna.
    - Contohnya persepsi konsumen terhadap botol air minum dalam kemasan, kebiasaan belanja (jarang, kadang-kadang, sering)



# Jenis Data

---

- **Berdasarkan sifat data**

- **Data Diskrit**

- data yang nilainya adalah bilangan asli.
    - Contoh : nilai mata uang rupiah dari waktu ke waktu

- **Data Kontinu**

- data yang nilainya ada pada suatu interval tertentu atau berada pada nilai yang satu ke nilai yang lainnya.
    - Contohnya penggunaan kata sekitar, kurang lebih, kira-kira dan sebagainya

# Jenis Data

---

- **Menurut waktu pengumpulannya**
  - **Data Cross Section**
    - data yang menunjukkan titik waktu tertentu
    - Contohnya laporan keuangan per 31 Desember 2006, data pelanggan PT. PLN bulan April 2006
  - **Data Time Series (Berkala)**
    - data yang datanya menggambarkan sesuatu dari waktu ke waktu atau periode secara historis
    - Contoh : nilai tukar rupiah terhadap dollar amerika dari tahun 2005 sampai 2006



# Sumber Data

## Internal Sources

- Databases: dapat dikuerikan via SQL, dll
- Teks dokumen
- Multimedia dokumen (audio, video)
- Spreadsheet (Excel, csv, JSON, dll)

## External Sources

- Open data repositories
- Public domain web pages

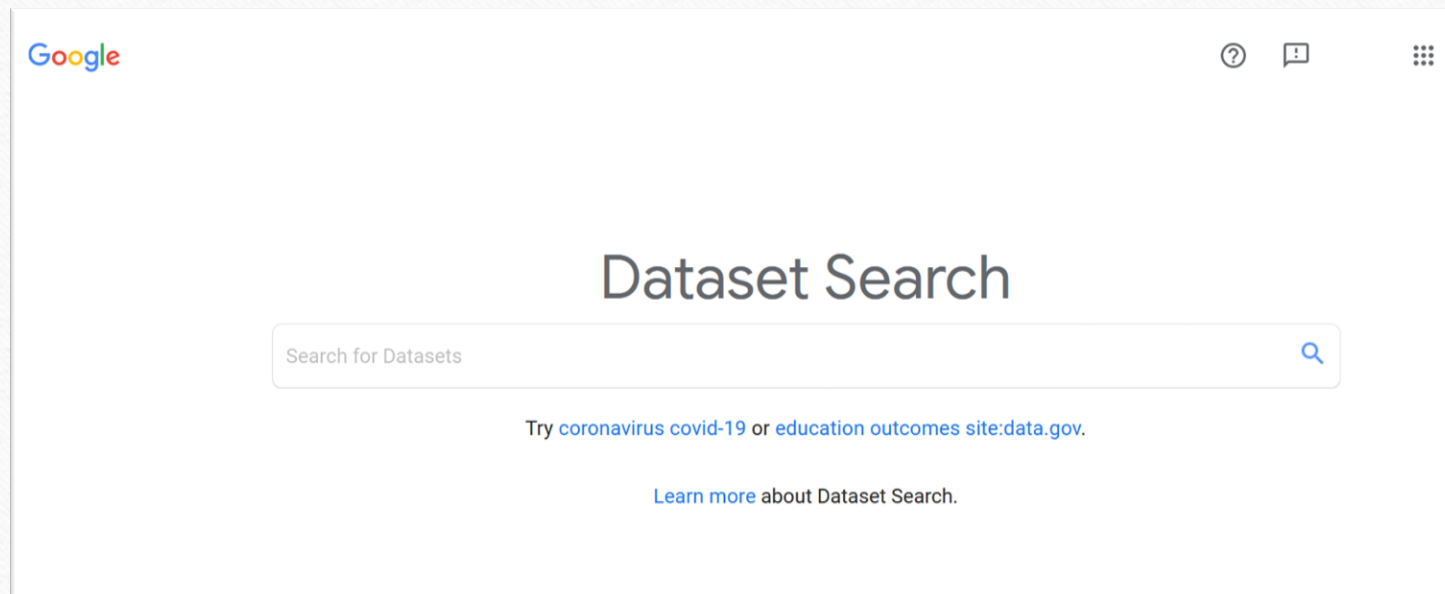
# Sumber data daring

- Portal Satu Data Indonesia (<https://data.go.id>)
- Portal Data Jakarta (<https://data.jakarta.go.id>)
- Portal Data Bandung (<http://data.bandung.go.id>)
- Badan Pusat Statistik (<https://www.bps.go.id>)
- Badan Informasi Geospasial (<https://tanahair.indonesia.go.id/>)
- UCI Machine Learning repository (<https://archive.ics.uci.edu/ml/index.php>)
- Kaggle (<https://www.kaggle.com/datasets>)
- World Bank Open Data (<https://data.worldbank.org>)
- UNICEF Data (<https://data.unicef.org>)
- WHO Open Data (<https://www.who.int/data>)
- IBM Data Asset eXchange (<https://developer.ibm.com/exchanges/data/>)
- DBPedia (<https://www.dbpedia.org/resources/>)
- Wikidata (<https://www.wikidata.org/>) .



# Sumber Data Daring

- Dapat juga dicari melalui google :  
<https://datasetsearch.research.google.com/>



# Kualitas Data

---

- **Accuracy:** benar atau salah, akurat atau tidak
- **Completeness:** ada yang tidak tercatat, tidak tersedia, ...
- **Consistency:** konsisten atau tidak
- **Timeliness:** apakah terupdate?
- **Believability:** seberapa dipercaya data itu benar?
- **Interpretability:** seberapa mudah data dapat dipahami?



# Kualitas Data

---

- Dipengaruhi oleh adanya :
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Kualitas alat pengambil data / alat ukurnya

# Data Preprocessing

---



## Data cleaning

Imputasi (missing values), smoothing (noisy data), identifikasi atau penghapusan (outliers), dan penanganan (data inconsistencies)



## Data integration

Integrasi dari multiple databases, data cubes, atau files



## Data reduction

Dimensionality reduction  
Numerosity reduction  
Data compression



## Data transformation and data discretization

Normalisasi  
Diskritasi data



# Problem Imbalanced Data

---

- Masalah Imbalanced Data perlu menjadi perhatian pada saat dilakukan penelitian pengklasifikasian. (Mengklasifikasikan suatu objek dalam kelas tertentu)
- Imbalanced Data dapat mengakibatkan overfitting
- Ada beberapa metode mengatasi Imbalanced Data tergantung dari domain data yang digunakan (numerik, teks, image, speech)

# Metode Pengumpulan Data

---

- Cara atau teknik yang dapat digunakan oleh peneliti untuk mengumpulkan data
- Metode yang digunakan :
  - Angket
  - Pengamatan
  - Wawancara
  - Tes
  - Analisis Dokumen, dll



# Instrumen Pengumpulan Data

---

- Alat bantu yang dipilih dan digunakan oleh peneliti dalam mengumpulkan data agar tindakannya sistematis dan lebih mudah.
- Contohnya :
  - Angket
  - daftar cek
  - Skala
  - pedoman wawancara
  - lembar / panduan pengamatan
  - soal ujian

# Kaitan Metode dan Instrumen

No	Jenis Metode	Jenis Instrumen
1	Angket (Kuisisioner)	Angket, Daftar cek (Checklist), Skala
2	Wawancara (Interview)	Pedoman Wawancara, Daftar cek
3	Pengamatan (Observasi)	Lembar pengamatan, Panduan pengamatan, Daftar cek
4	Ujian atau Tes	Soal Ujian /'Tes
5	Dokumentasi	Daftar cek, Tabel



# Angket

---

- Suatu teknik atau cara pengumpulan data secara tidak langsung (peneliti tidak langsung bertanya jawab dengan responden)
- Jenis Angket :
  - Angket tertutup
    - Keuntungan : mudah diolah, responden tidak perlu menuliskan buah pikirannya, waktu pengisian relatif singkat, dapat menjangkau responden yang relatif banyak
  - Angket terbuka
    - Keuntungan : responden dapat mengungkapkan buah pikirannya, berguna jika peneliti ingin mengetahui keadaan responden lebih mendalam

# Wawancara

---

- Cara pengumpulan data yang digunakan untuk memperoleh informasi langsung dari sumbernya
- Digunakan bila ingin mengetahui hal hal dari responden yang jumlahnya sedikit secara mendalam
- Faktor yang mempengaruhi arus informasi dalam wawancara adalah pewawancara, responden, pedoman wawancara dan situasi wawancara



# Pengamatan

---

- Suatu teknik atau cara untuk mengumpulkan data dengan jalan mengamati kegiatan yang sedang berlangsung
- Ada 2 jenis :
  - Pengamatan Partisipatori :
    - Pengamat ikut serta dalam kegiatan yang sedang berlangsung
    - Contoh : pengamat ikut serta dalam rapat atau pelatihan
  - Pengamatan Nonpartisipatori :
    - Pengamat tidak ikut serta dalam kegiatan yang berlangsung, hanya berperan mengamati kegiatan

# Tes

---

- Adalah suatu prosedur yang sistematis untuk mengamati atau mendeskripsikan satu atau lebih karakteristik seseorang dengan menggunakan standar numerik atau sistem kategori
- Contoh :
  - Tes kepribadian
  - Tes bakat
  - Tes prestasi (achievement test)
  - Tes intelegensi
  - Tes Sikap



# Dokumentasi

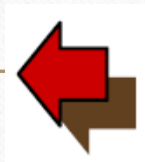
---

- Ditujukan untuk memperoleh data langsung dari tempat penelitian, meliputi buku, peraturan, laporan kegiatan, foto, film documenter dan data yang relevan dengan penelitian.

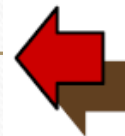
# Ilustrasi Alur Data



Proses Data



Preprocessing  
Data



Pengumpulan  
Data

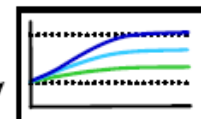


Hasil Pengujian

**Data** antar tabel



Analisa Hasil



Kesimpulan :

.....  
.....

Tahun	2010	2011	2012	2013	2014	2015	2016	2017	2018
Produk A	100	120	150	180	200	220	250	280	300
Produk B	80	90	100	110	120	130	140	150	160
Produk C	60	70	80	90	100	110	120	130	140
Produk D	40	50	60	70	80	90	100	110	120

Kesimpulan :

.....  
.....



# Pengertian Variabel

---

- Variabel penelitian :
  - Atribut atau sifat, atau nilai dari orang, objek atau kegiatan yang mempunyai variasi tertentu yang ditetapkan oleh peneliti untuk dipelajari, kemudian ditarik kesimpulannya
- Ciri variabel penelitian :
  - Memiliki variasi nilai
  - Membedakan satu objek dengan objek lain dalam satu populasi
  - Dapat diukur

# Macam-macam Variabel berdasarkan fungsinya

---

- Variabel Bebas:
  - Mempengaruhi variable terikat, baik pengaruh positif ataupun negative
  - Disebut juga variable independen, variable predictor atau variable anteseden
- Variabel Terikat
  - Variabel yang dijelaskan atau dipengaruhi oleh variable bebas
  - Disebut juga variable dependen, variable keluaran (output), variable kriteria atau variable konsekuen
- Misal : Pengaruh **resolusi citra** terhadap **akurasi klasifikasi**, mana variable bebas dan terikatnya?



# Lanjutan

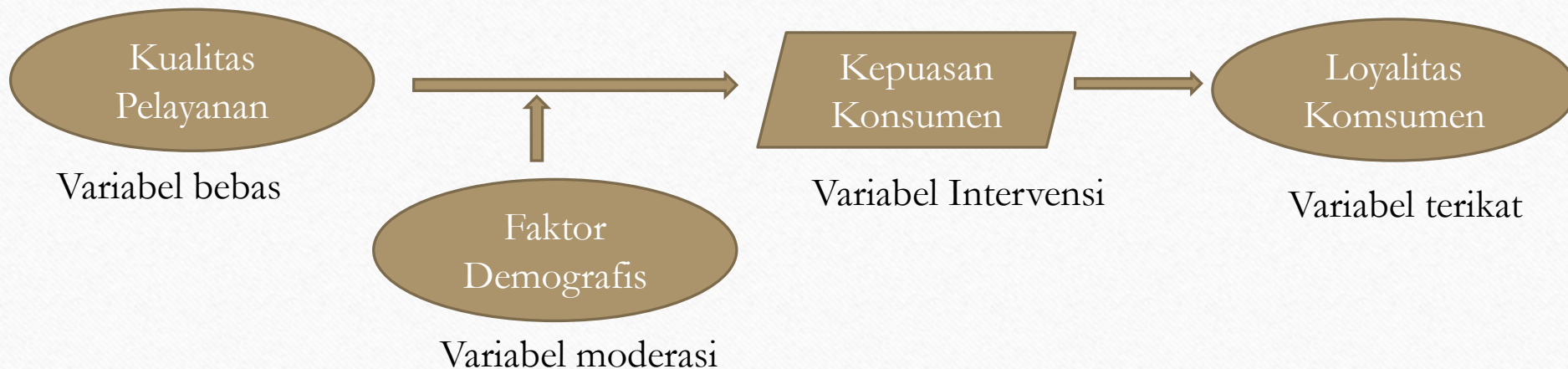
---

- Variabel Moderasi
  - Variabel yang mempengaruhi (memperkuat atau memperlemah) hubungan antara variable bebas dan variable terikat.
  - Contoh :
    - Hasil belajar mahasiswa dipengaruhi motivasi belajar
    - Sikap dosen dapat dijadikan variable moderasinya
    - Sikap dosen tegas ➔ Sikap positif, memotivasi mahasiswa
    - Sikap dosen arogan ➔ sikap negative, menurunkan motivasi belajar mahasiswa

# Lanjutan

- Variabel Intervensi

- Mempengaruhi hubungan langsung antara variable bebas dan variable terikat sehingga terjadi hubungan tidak langsung.
- Biasanya terdapat dalam penelitian yang lebih kompleks





# Validitas

---

- Konsep yang berkaitan dengan sejauh mana tes telah mengukur apa yang seharusnya diukur
- Ada dua jenis :
  - Validitas logis
  - Validitas empiris

# Validitas Logis

---

- Untuk menentukan berfungsi tidaknya soal berdasarkan kriteria yang telah ditentukan, yang dalam hal ini kriteria materi, konstruksi dan bahasa.
- Suatu tes atau instrument pengukuran bervaliditas tinggi apabila alat tersebut menjalankan fungsi ukurnya atau memberikan hasil ukur yang sesuai dengan maksud dilakukannya pengukuran tersebut
- Jika peneliti menggunakan kuesioner di dalam pengumpulan data penelitian, maka item-item yang disusun pada kuesioner tersebut merupakan alat test yang harus mengukur apa yang menjadi tujuan penelitian.



# Validitas Empiris

---

- Validitas yang ditentukan berdasarkan kriteria baik kriteria internal (kriteria tes atau instrument itu sendiri) maupun kriteria eksternal (kriteria hasil ukur tes/ diluar instrument)

# Validitas dan Reliabilitas

---

- Suatu skala pengukuran dikatakan valid apabila skala tersebut digunakan untuk mengukur apa yang seharusnya diukur.
- Misalnya :
  - skala nominal (bersifat non- parametrik) digunakan untuk mengukur variabel nominal bukan untuk mengukur variabel interval (bersifat parametrik).
- Sesuatu dikatakan valid jika alat ukur yang dibuat sesuai dengan apa yang hendak diukur,
- Misal :
  - jika yang diukur adalah panjang, maka penggaris adalah alat ukur yang valid.

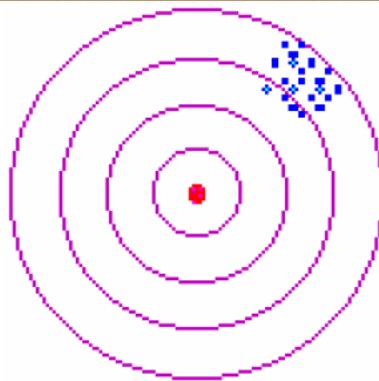


# Validitas dan Reliabilitas

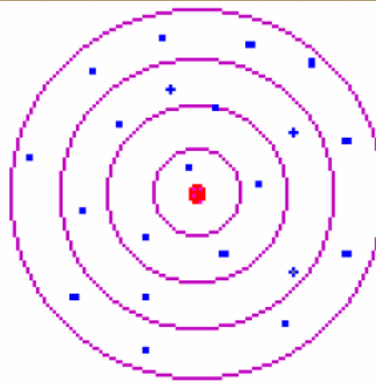
---

- Reliabilitas atau keterandalan
  - mengukur sesuatu secara konsisten, apapun yang diukur dan jika pengukuran dilakukan dalam kondisi apapun akan memberikan hasil yang sama dari data yang dikumpulkan.
- suatu alat ukur yang tidak reliable pasti tidak valid
- alat ukur yang reliable belum tentu valid

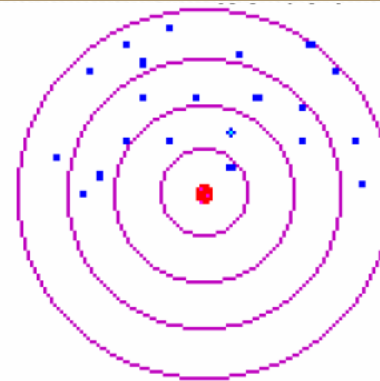
# Validitas dan Reliabilitas



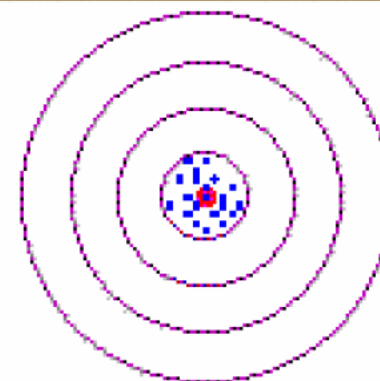
**Reliable  
Not Valid**



**Valid  
Not Reliable**



**Neither Reliable  
Nor Valid**



**Both Reliable  
And Valid**



# Referensi

---

- Chastine Fatichah, Pengertian Data, Jenis dan Tipe Data, Deskripsi Data, Data Preprocessing, 2021
- Zainal A. Hasibuan, Metodologi Penelitian pada Bidang Ilmu Komputer dan Teknologi Informasi
- Andi Wahyu Rahardjo E, Petunjuk Praktis Metode Penelitian Teknologi Informasi