**Rangkuman Data untuk AI**

Kita sudah berada di penghujung materi Data untuk AI. Sampai sini, Anda telah memiliki pemahaman mendasar mengenai data untuk AI. Mari kita rangkum secara saksama.

**Pengenalan Data**

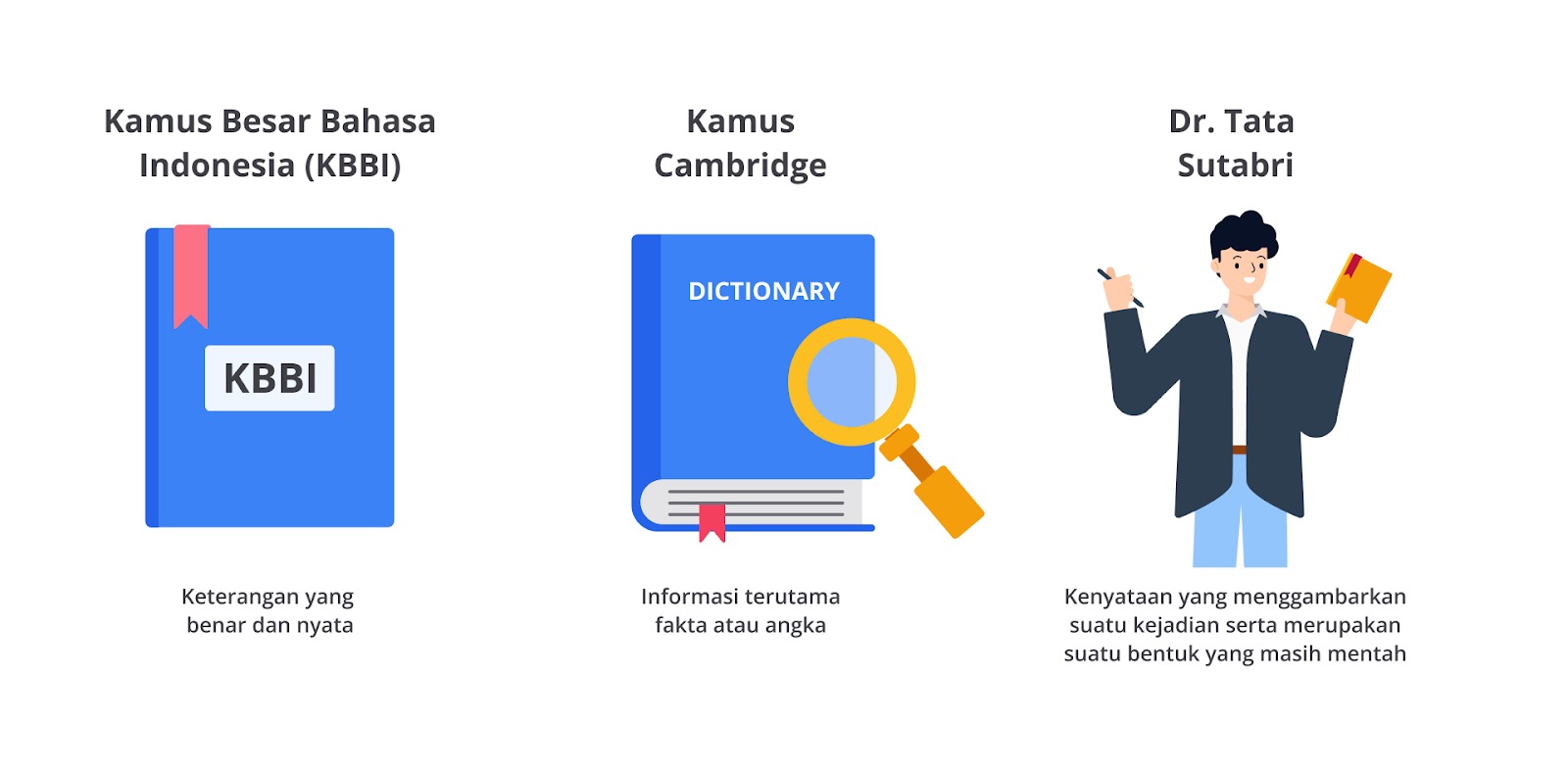
**Apa Itu Data**

Data merupakan fondasi penting dalam pengembangan AI. Kualitas dan jumlah data yang baik dapat memengaruhi hasil pengembangan AI menjadi lebih baik. Setelah mengetahui bahwa data memiliki peran penting dalam proses pengembangan AI mari kita gali sedikit lebih dalam tentang data. Mungkin pertanyaan pertama yang terbesit di benak Anda saat ini, “Apa sih data itu?” Yuk, kita bahas satu per satu.

Menurut Kamus Besar Bahasa Indonesia, data merupakan keterangan yang benar dan nyata; keterangan atau bahan nyata yang dapat dijadikan dasar kajian; informasi dalam bentuk yang dapat diproses oleh komputer, seperti representasi digital dari teks, angka, gambar grafis, atau suara.

Menurut Kamus Cambridge, data adalah informasi, terutama fakta atau angka, dikumpulkan untuk diperiksa dan dipertimbangkan, serta digunakan untuk membantu pengambilan keputusan atau informasi dalam bentuk elektronik yang dapat disimpan dan digunakan oleh komputer.

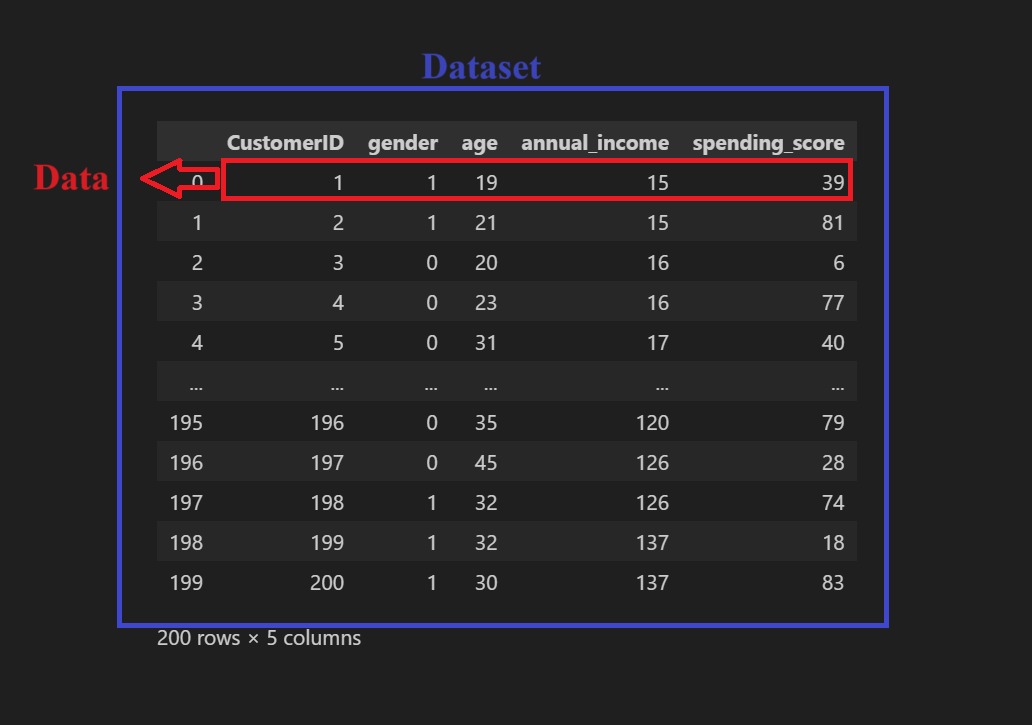
Terakhir, menurut ahli, yaitu Tata Sutabri, dalam buku Konsep Sistem Informasi, data adalah kenyataan untuk menggambarkan suatu kejadian serta suatu bentuk yang masih mentah dan belum dapat bercerita banyak sehingga perlu diolah lebih lanjut melalui suatu model untuk menghasilkan informasi.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

**Data, Dataset, dan Basis Data**

Pemahaman yang baik tentang perbedaan antara data, dataset, dan basis data memungkinkan Anda mengelola dan memanfaatkan informasi dengan lebih efisien, membuat keputusan yang lebih tepat, dan menghindari kebingungan atau kesalahan yang mungkin terjadi dalam konteks pengembangan AI.

Data adalah fakta, nyata, dan informasi yang tersimpan di dalamnya dapat berbentuk teks, angka, gambar, suara, dan banyak bentuk lainnya. Data dalam konteks dataset dan basis data mengacu pada kumpulan informasi yang relevan serta dikumpulkan, disimpan, dan dikelola untuk tujuan tertentu. Dengan kata lain data merupakan entri tunggal atau informasi individual. Perhatikan gambar berikut.

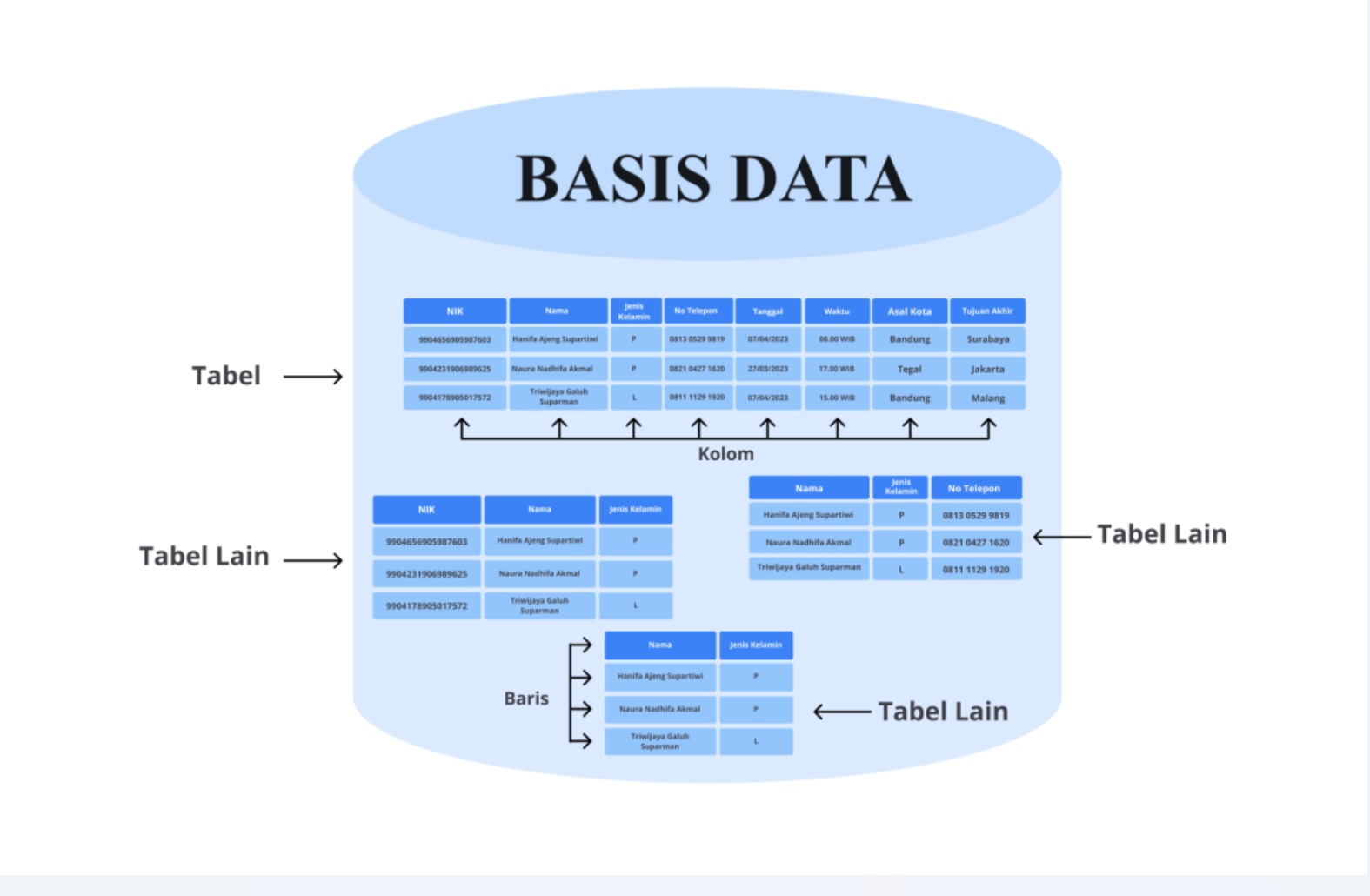
[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Setelah melihat gambar di atas, dapatkah Anda mengamati perbedaan antara data dan dataset? *Yup!* Dataset adalah kumpulan data yang disusun secara terstruktur. Dataset dapat digunakan untuk tujuan tertentu, seperti pembangunan machine learning, analisis statistik, dan visualisasi data. Lalu pada contoh di atas, yang mana data? Seperti yang telah dibahas sebelumnya, data merupakan entri tunggal. Pada kasus di atas, data mencakup satu baris dari dataset yang ada dan dapat disebut sebagai data pelanggan.

Selain data dan dataset, Anda juga perlu mengetahui tentang basis data karena keduanya memiliki hubungan yang erat. Mari kita mulai menyelam sedikit lebih dalam agar dapat mengetahui perbedaan di antara keduanya!

Basis Data merupakan kumpulan data yang diatur dan disimpan secara terorganisir sehingga dapat diambil dan diakses dengan mudah. Selain itu, ia juga dapat menyimpan berbagai macam tipe data, termasuk teks, nomor, gambar, dan tipe data lainnya. Lalu, apa perbedaannya dengan dataset? Basis data memiliki banyak kumpulan data dan dapat digunakan untuk aplikasi yang berbeda, sedangkan dataset merupakan bagian dari data yang diambil dari basis data.

Agar Anda dapat lebih memahami perbedaan antara dataset dan basis data, yuk telisik bersama gambar di bawah ini.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Gambar tabung di atas merupakan representasi dari sebuah basis data dan tabel yang di dalamnya memiliki struktur basis data yang diatur dengan kolom dan baris. Namun, berdasarkan gambar di atas, apakah Anda dapat menentukan manakah yang termasuk dataset? Benar, Jika tabel yang ada dalam basis data dibungkus dengan menggunakan format tertentu dan dibuat menjadi satu buah tabel maka ia dapat disebut dataset.

**Tipe-Tipe Data**

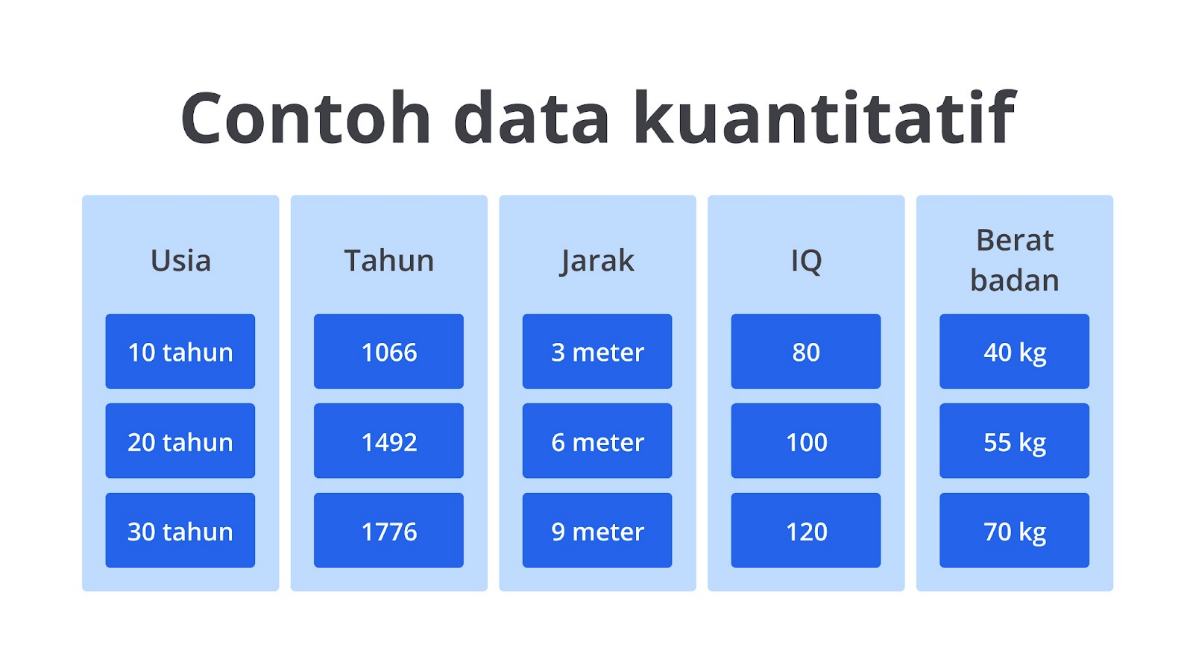
Seperti halnya manusia yang memiliki tipenya tersendiri, pengembangan AI juga memerlukan data yang sesuai dengan “tipe”-nya agar dapat mengerjakan tugasnya dengan baik. Oleh karena itu, pada modul ini, kita akan bersama-sama mengetahui berbagai macam tipe data mulai dari data terstruktur hingga data tidak terstruktur.

**Data Terstruktur**

Data terstruktur merupakan jenis data yang memiliki format dan tata letak yang tetap atau teratur. Artinya, data ini diatur dalam suatu pola atau struktur yang konsisten sehingga mudah dibaca, diproses, dan dianalisis oleh komputer atau manusia. Jenis data terstruktur umumnya memiliki definisi yang jelas seperti kolom dalam tabel atau bidang dalam dokumen teks. Data ini memiliki 2 turunan, yaitu data kuantitatif dan data kategorikal.

**Data Kuantitatif**

Data kuantitatif adalah jenis data yang dapat diukur atau diungkapkan dalam bentuk angka. Data ini digunakan untuk mengukur atau menggambarkan jumlah, besaran, atau atribut-atribut yang dapat diukur secara numerik. Perhatikan contoh data kuantitatif berikut.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Pada modul ini, kita akan menyelam lebih dalam karena data kuantitatif akan terbagi menjadi dua bagian, yaitu data kontinu dan diskrit. Mari kita kupas tuntas hingga ke akarnya!

***Data Kontinu***

Data kontinu dapat direpresentasikan dalam berbagai nilai numerik, seperti bilangan desimal, bulat, dan lain-lain. Beberapa contoh tipe data kontinu yang umum adalah tinggi, berat, waktu, suhu, usia, dan lain-lain.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Mari kita analogikan bahwa Anda memiliki sebuah toko kelontong dan ingin menentukan data kontinu pada kasus yang Anda miliki. Sebelum Anda membuka toko, ada baiknya jika mengetahui suhu yang terjadi di toko agar Anda dapat memutuskan ingin menyalakan AC pada waktu yang tepat sehingga pengunjung merasa nyaman. Sehingga Anda akan mencatat data sebagai berikut.

| **Waktu** | **Suhu (Celcius)** |
| --- | --- |
| 10.00 | 24 |
| 10.15 | 24.5 |
| 10.30 | 24.75 |
| … | … |

Dari data tersebut, Anda dapat menentukan waktu penggunaan *AC* yang tepat supaya pengunjung merasa nyaman. Misalnya, ketika suhu sudah lebih dari 28C, AC harus dinyalakan.

***Data Diskrit***

Data diskrit merupakan data numerik yang hanya bisa direpresentasikan dengan bilangan bulat dan tidak dapat dibagi ke dalam unit yang lebih kecil. Perhatikan tabel berikut.

| **Produk** | **Stok** |
| --- | --- |
| Kopi Brazil | 12 |
| Susu THT | 15 |
| Snack Cheetos | 7 |
| Beras | 5 |

Tabel tersebut menunjukan stok produk yang sudah mulai habis pada hari itu. Karena stok produk merupakan sebuah objek tunggal, ia tidak dapat dibagi ke dalam unit yang lebih kecil. Oleh karena itu, stok produk termasuk ke dalam data diskrit.

**Data Kategorikal**

Data kategorikal mengacu pada bentuk informasi yang dapat disimpan dan diidentifikasi berdasarkan nama atau labelnya. Data kategorikal merupakan data yang dapat dikelompokkan dan terbagi berdasarkan karakteristik atau ciri khasnya masing-masing. Dari data kategorikal, ada dua pembagian, yaitu ordinal dan nominal.

***Data Ordinal***

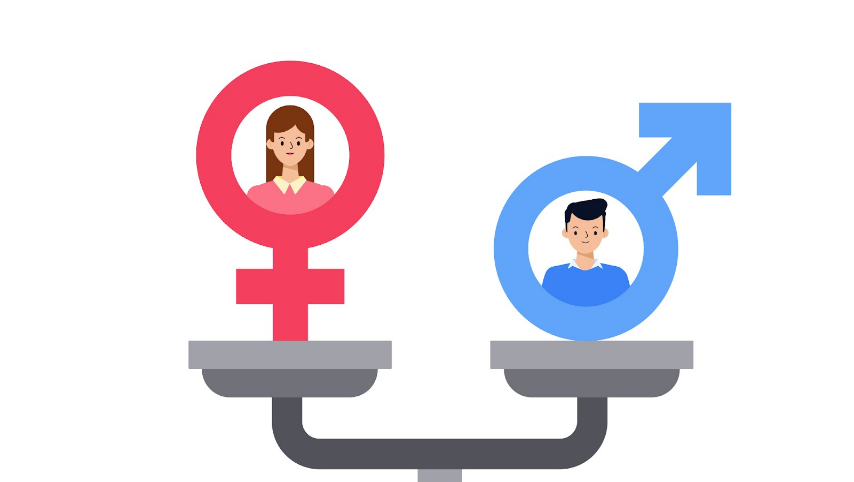
Data ordinal adalah jenis pengelompokan data yang memiliki urutan atau harus disusun secara berurutan dengan mekanisme peringkat. Perhatikan data rating yang ada pada gambar berikut.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Gambar di atas menunjukkan bahwa terdapat dua kelas yang memiliki rating 4,86 dan 4,83. Dari kedua kelas tersebut, manakah menurut Anda yang lebih baik? Dari angka tersebut, pasti Anda langsung bisa mengetahui kelas manakah yang memiliki rating tertinggi. Nah informasi tersebut merupakan contoh dari data ordinal bahwa jenis data ini dapat diurutkan berdasarkan peringkat.

***Data Nominal***

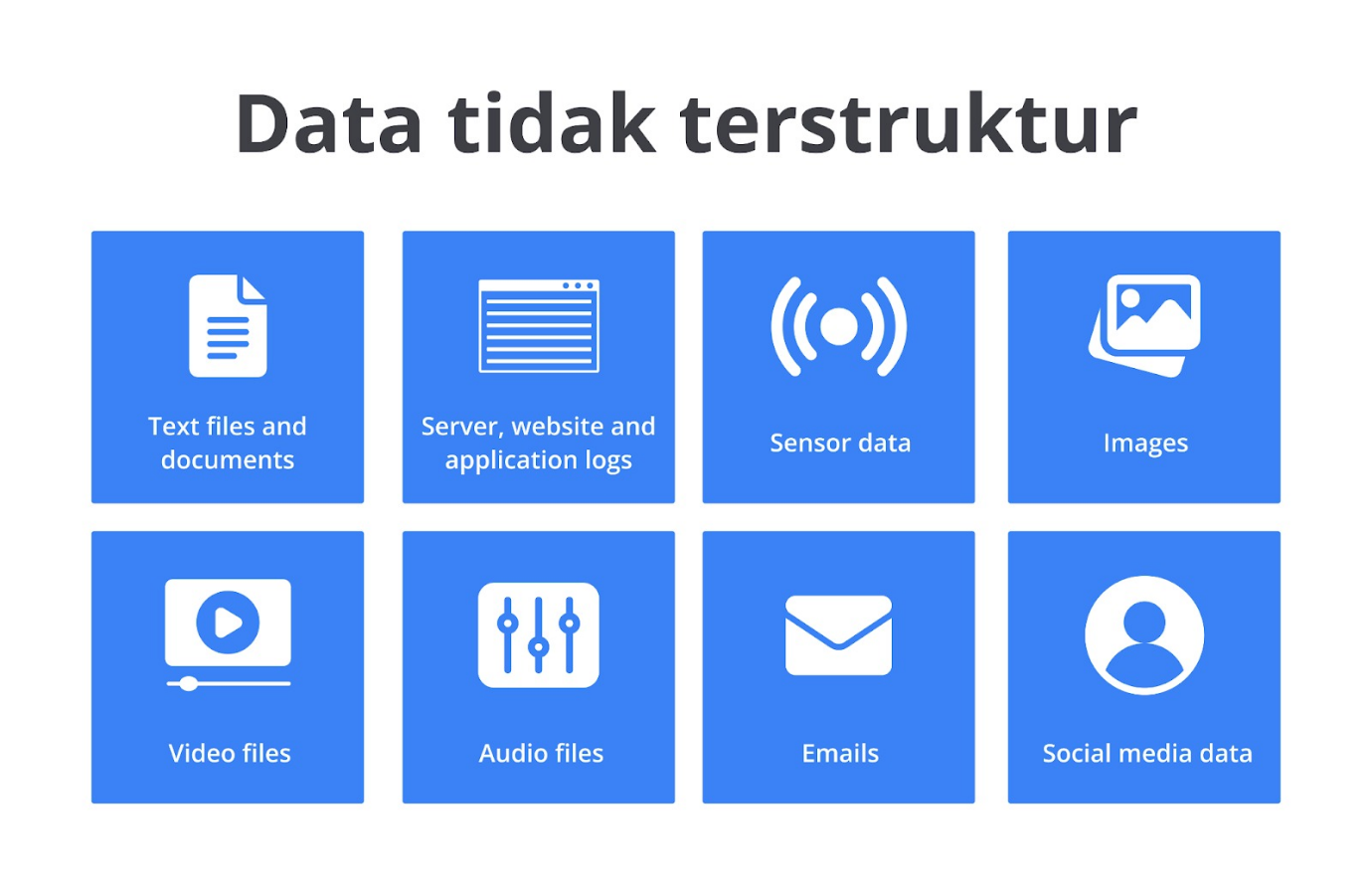
Berkebalikan dengan data ordinal, data nominal adalah jenis pengelompokan data yang tidak memiliki keterkaitan dengan data lainnya dan tidak memiliki arti khusus. Jadi, data ini dapat dibedakan tanpa harus mengurutkan atau dibandingkan dengan data lainnya.  Perhatikan gambar di bawah ini.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Seperti yang tertera pada gambar di atas, Anda dapat mengetahui bahwa jenis kelamin terbagi menjadi 2, yaitu perempuan dan laki-laki. Pada dasarnya jenis kelamin tersebut tidak memiliki keterkaitan, tetapi dapat diklasifikasikan menjadi jenis kelamin. Seperti halnya pengertian dari data nominal, data ini tidak dapat dibandingkan atau diurutkan.

**Data Tidak Terstruktur**

Data tidak terstruktur adalah jenis data yang tidak memiliki format atau struktur yang jelas. Data ini cenderung bervariasi bentuknya dan sulit untuk diorganisasi dalam kategori atau kolom tertentu. Data tidak terstruktur seringkali memiliki sifat lebih bebas, tidak terbatas, dan lebih kompleks dibandingkan dengan data terstruktur. Berikut merupakan contoh dari data tidak terstruktur.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

**Kriteria data untuk AI**

Kenapa kita harus mempelajari ini? Karena ada salah satu pepatah yang mengatakan “*garbage in, garbage out*” atau GIGO yang artinya bahwa output suatu sistem komputer tidak akan lebih baik dibandingkan inputnya. Maksudnya apa *sih*? Jika Anda penasaran dengan kalimat tersebut, berarti Anda berada di jalur yang benar. Yuk, kita pelajari bersama-sama detail dari GIGO!

**Garbage in, Garbage out (GIGO)**

*Garbage in, garbage out*yang dalam bahasa Indonesia memiliki arti sampah masuk, sampah keluar. Apakah Anda memahami arti kata “sampah” di sini? Mari kita samakan sudut pandang kita terkait kata tersebut supaya tidak terjadi salah paham.

Sampah di sini berarti data yang buruk, tidak wajar, tidak relevan, dan keliru sehingga nantinya akan menghasilkan AI yang tidak sesuai dengan harapan pengembang.

Perhatikan gambar berikut.

[](https://www.dicoding.com/academies/653/tutorials/35413?from=35408)

Dari gambar di atas kita bisa mendapatkan kesimpulan bahwa keluaran dari AI yang kita bangun sangat bergantung pada data masukkan yang diterima. Jika kita memiliki data masukkan yang buruk, besar kemungkinan AI yang dihasilkan tidak sesuai harapan.

Sedari tadi, kita telah membahas kualitas data dengan menyebut “data yang buruk”. Namun, hingga saat ini kita belum membahas apa saja masalah yang ada pada data sehingga sebuah data dapat memiliki kualitas yang buruk.

**Masalah dalam Data**

Sebelum kita memasuki permasalahan pada data, alangkah baiknya kita mengetahui terlebih dahulu mengapa bisa terjadi permasalahan dalam data. Permasalahan yang terjadi pada data biasanya disebabkan oleh kesalahan ketika pengumpulan atau pencatatan data. Data yang diperoleh dengan cara yang salah atau bahkan diambil dari sumber yang tidak dapat dipercaya juga bisa disebut sebagai data sampah.

Dari hal tersebut muncullah permasalahan umum pada data. Permasalahan umum yang sering terjadi terdapat pada kualitas data di antaranya, seperti data yang tidak relevan (*incorrect*), data berbeda dengan yang lain (*outlier*), data duplikat, data kosong, data yang tidak benar, dan masih banyak lagi. Pada kesempatan ini, kita akan membahas beberapa permasalahan umum yang terjadi dan bagaimana cara menanganinya.

Dari beberapa permasalahan di atas, setidaknya kita akan sering menemui data kosong. Kita dapat menanganinya dengan beberapa cara, seperti menghapus data yang kosong tersebut atau mengisi data kosong dengan nilai rata-rata atau median jika datanya berupa numerik.

**Infrastruktur Data di Industri**

Tujuan dari infrastruktur data adalah untuk menyediakan pengelolaan data yang baik, memproses data, dan menganalisis data yang ada. Jika kalian penasaran dengan tujuan infrastruktur data tersebut, mari kita jabarkan satu per satu tujuan dari infrastruktur data di industri.

1. **Manajemen data**  
   Dengan menggunakan infrastruktur data yang baik maka tempat penyimpanan data akan terpusat. Hal ini akan membuat data dalam sebuah organisasi menjadi lebih aman dan mudah untuk dikelola.
2. **Pemrosesan data**  
   Infrastruktur data menyediakan daya komputasi dan sumber daya yang dibutuhkan untuk memproses dan menganalisis data dengan jumlah besar. Hal ini memungkinkan organisasi untuk melakukan analisis dan membuat pemodelan yang kompleks sehingga dapat membantu untuk mendapatkan informasi dan keputusan yang tepat berdasarkan data.
3. **Integrasi data**  
   Seperti yang sudah dijelaskan sebelumnya, dengan menggunakan infrastruktur data yang baik, kita dapat mengintegrasikan data dari berbagai sumber.
4. **Keamanan data**  
   Infrastruktur data menyediakan fitur dan protokol keamanan untuk melindungi data sensitif dari akses yang tidak sah, pencurian, atau penyalahgunaan. Hal ini memastikan kepatuhan terhadap peraturan dan praktik terbaik untuk keamanan dan privasi data.