**Pengenalan Data**

Pada modul ini Anda akan belajar mengenal data secara umum, kriteria data untuk AI, hingga gambaran infrastruktur data di industri agar dapat memahami pentingnya data dalam proses pengembangan AI.

Materi yang akan dibahas pada modul ini antara lain sebagai berikut.

* Pengertian data
* Perbedaan data, dataset, dan basis data
* Gambaran infrastruktur data di industri

Mari kita mulai pembelajaran terkait data untuk AI agar Anda lebih dekat kepada pengembangan AI yang sesungguhnya. Semangat!

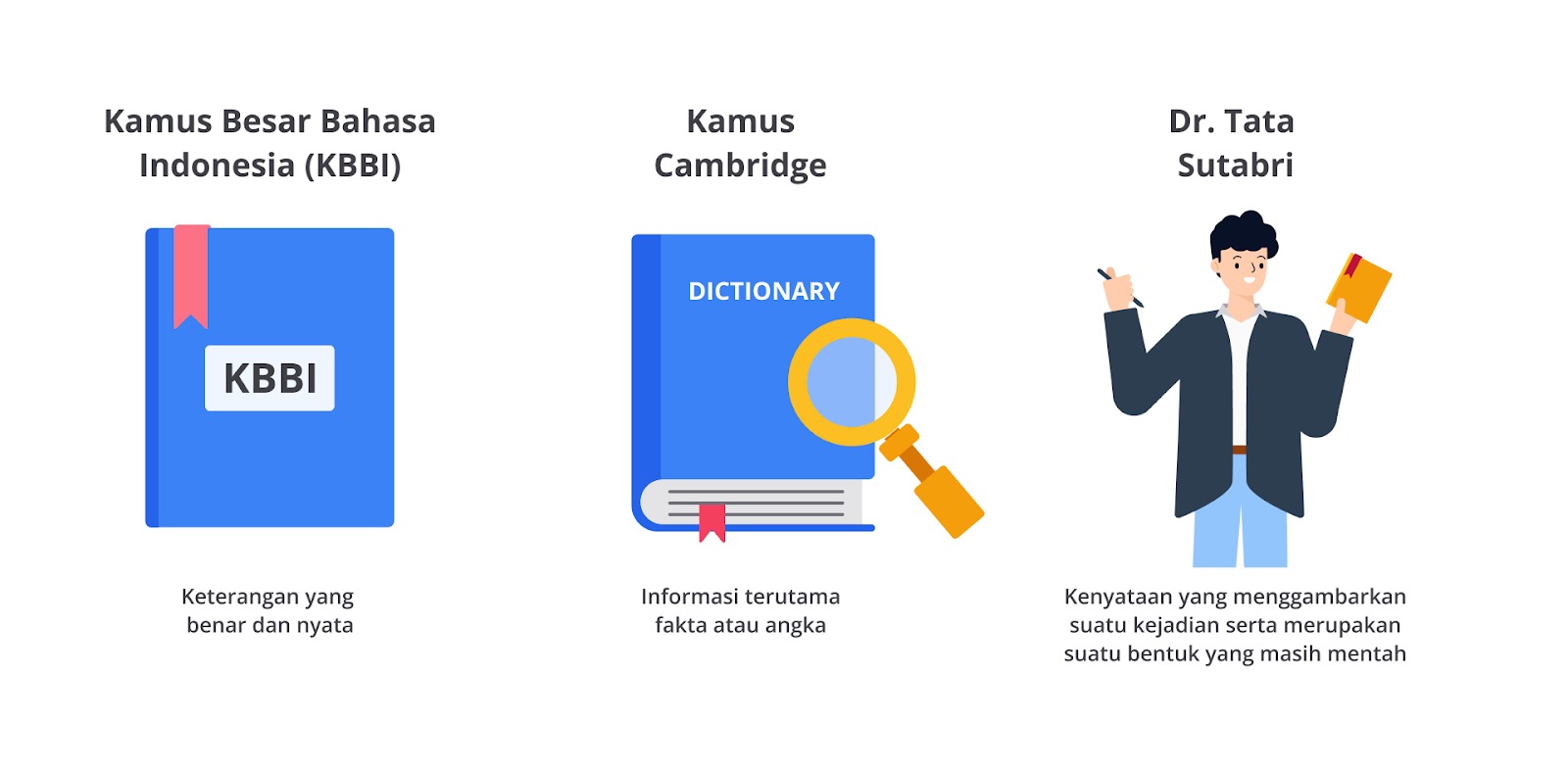
**Apa Itu Data**

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.



Dari ketiga pengertian di atas, dapat disimpulkan arti data, yaitu Informasi yang benar, fakta, dan nyata. Dengan kata lain, data merupakan kenyataan apa adanya (raw facts).

Bagaimana menurut Anda? Apakah Anda sudah memahami tentang data? Supaya pemahaman Anda lebih lengkap, selanjutnya kita akan menelaah perbedaan data, dataset, dan basis data karena ketiga hal tersebut memiliki kaitan satu sama lain.

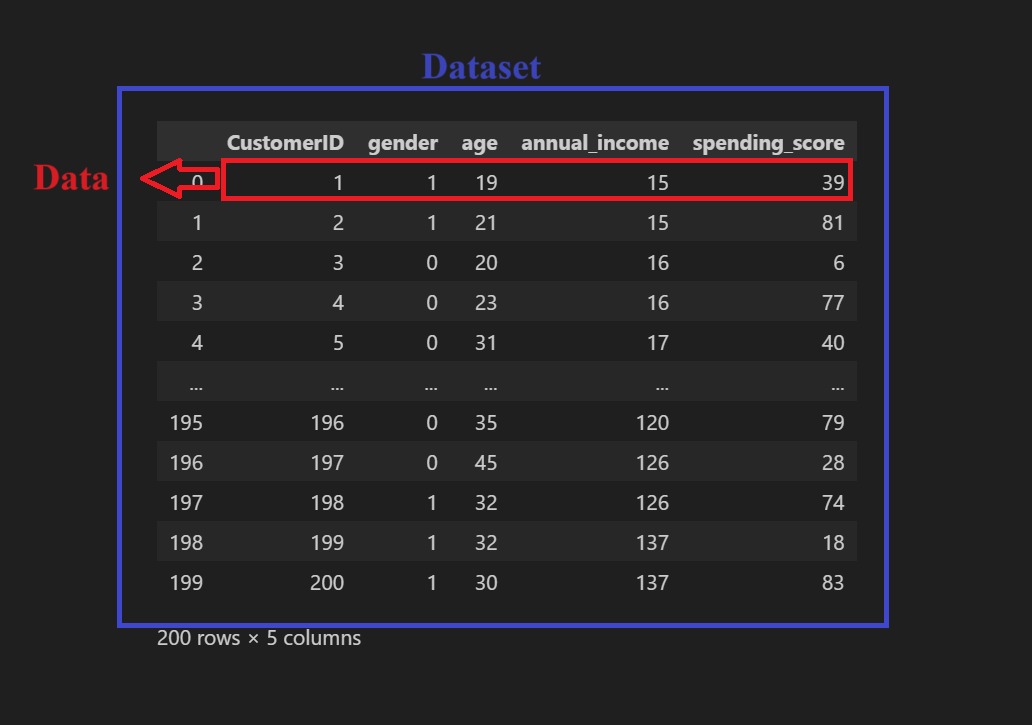
**Data, Dataset, dan Basis Data**

Pada bagian ini, Anda akan mengidentifikasi perbedaan antara ketiga hal yaitu data, basis data, dan dataset. Lalu, mengapa kita harus mengetahui perbedaan dari ketiga materi tersebut? Pertanyaan yang bagus!

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.

Untuk mengetahui perbedaan dasar dari data, dataset, dan basis data, mari kita mengingat kembali dari awal, yaitu data.

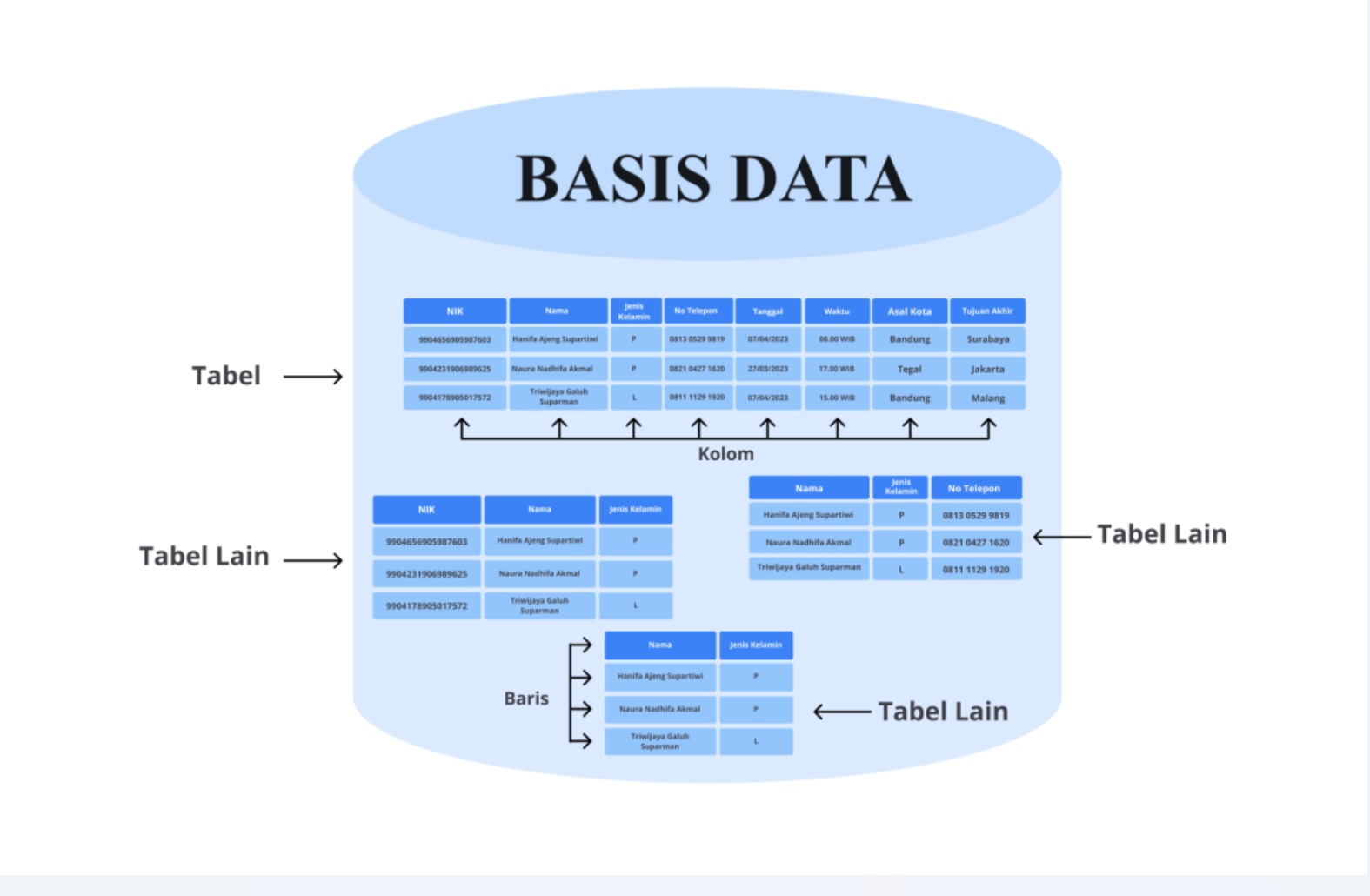
Data adalah fakta, nyata, dan informasi yang tersimpan di dalamnya dapat berbentuk teks, angka, gambar, suara, dan banyak bentuk lainnya. Data dalam konteks dataset dan basis data mengacu pada kumpulan informasi yang relevan serta dikumpulkan, disimpan, dan dikelola untuk tujuan tertentu. Dengan kata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



Setelah melihat gambar di atas, dapatkah Anda mengamati perbedaan antara data dan dataset? *Yup!* Dataset adalah kumpulan data yang disusun secara terstruktur. Biasanya, dataset dipresentasikan dalam bentuk tabel alias kumpulan baris dan kolom yang dapat disimpan pada beberapa format, seperti CSV, Excel, JSON, dan format lainnya. 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.

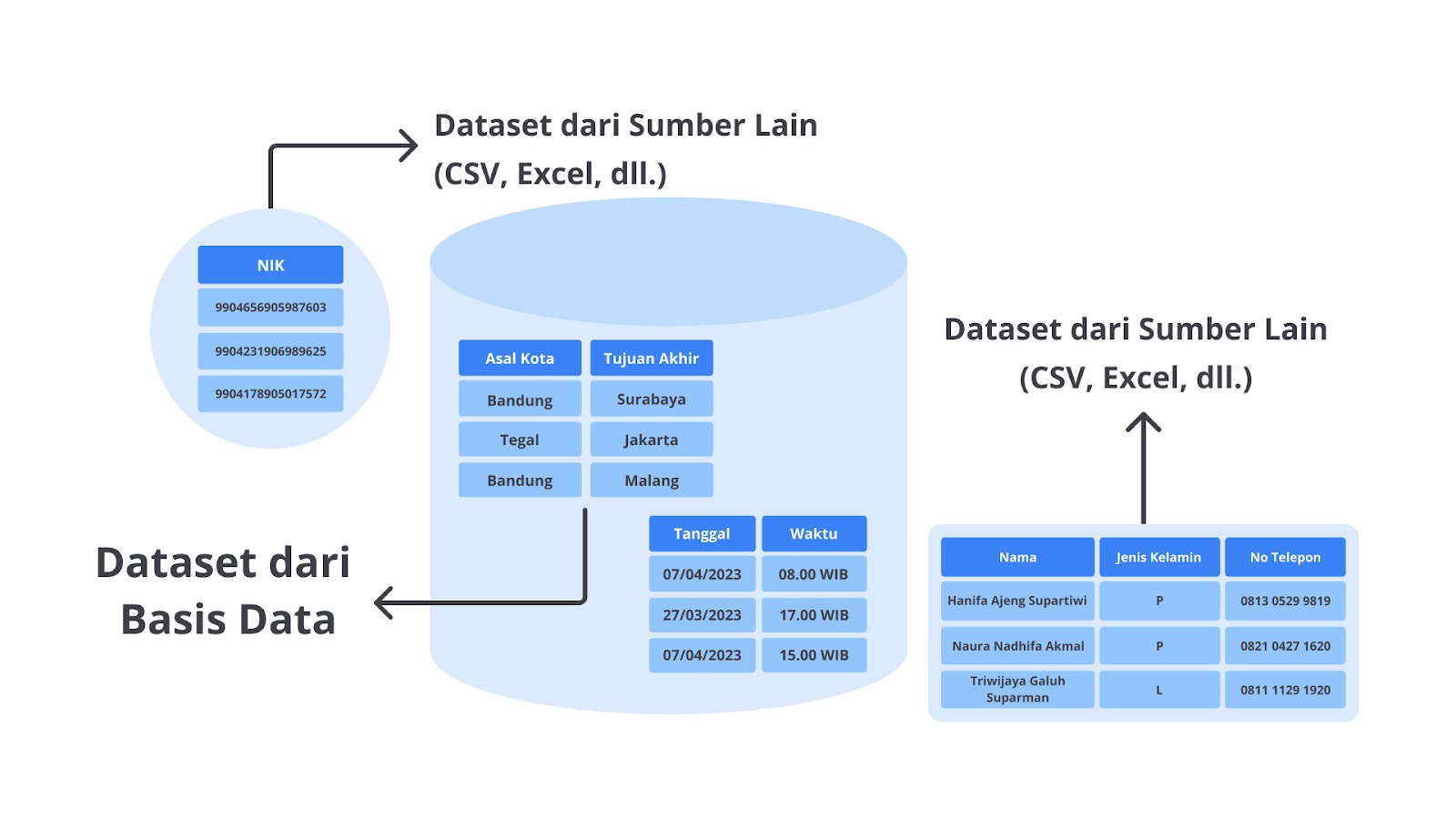
Selain data dan dataset, Anda juga perlu mengetahui tentang basis data karena keduanya memiliki hubungan yang erat. Mari kita mulai menyelam sedikit lebih dalam agar dapat mengetahui perbedaan di antara keduanya!

**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 kumpulan dari data yang diambil dari basis data.



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.

Jika kalian ingat dataset merupakan kumpulan data yang disusun secara terstruktur, lalu dari manakah dataset diperoleh? Dataset bisa kita dapatkan baik dari basis data maupun dari sumber yang relevan dengan objek yang sedang dikerjakan. Berikut beberapa contoh dari pengumpulan dataset.



**Tipe-Tipe Data**

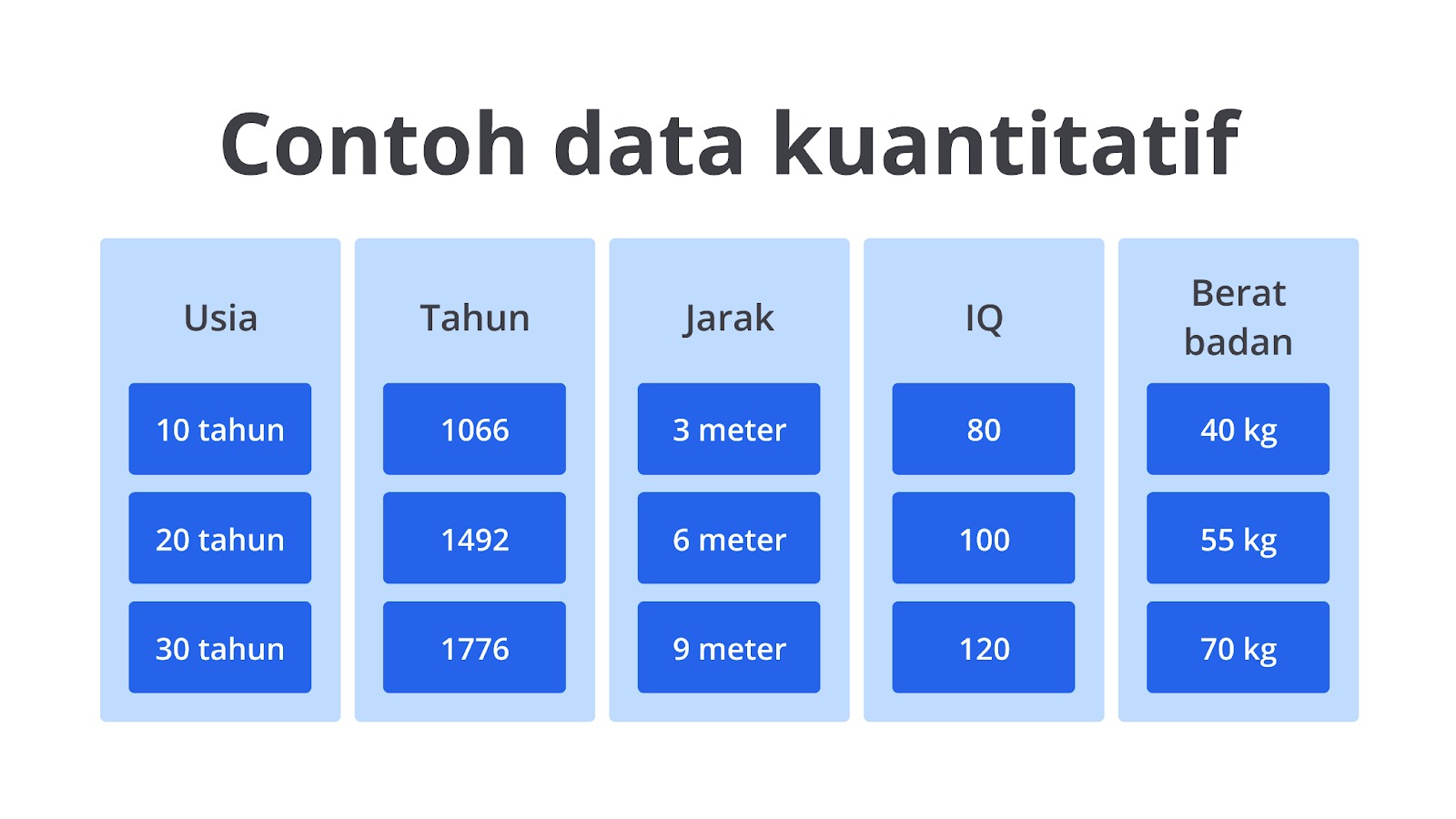
*Halohaa!* Pada materi sebelumnya, Anda sudah menelaah perbedaan data, dataset, dan basis data. Nah, sekarang saatnya mempelajari tipe-tipe pada data. Sebenarnya, untuk apa kita mempelajari tipe data? Seperti halnya manusia yang memiliki tipenya tersendiri, pengembangan AI juga memerlukan data yang sesuai 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. Yuk, berangkat!

**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. Mari kita bahas keduanya!

**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.



***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.

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 perubahan temperatur yang terjadi di toko agar Anda dapat memutuskan ingin menyalakan AC pada waktu yang tepat sehingga pengunjung merasa nyaman. Anda dapat menentukan waktu penggunaan *AC* yang tepat supaya pengunjung merasa nyaman. Misalnya, ketika suhu sudah lebih dari 28º C, 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. 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. Penasaran? Yuk, kita bahas!

***Data Ordinal***

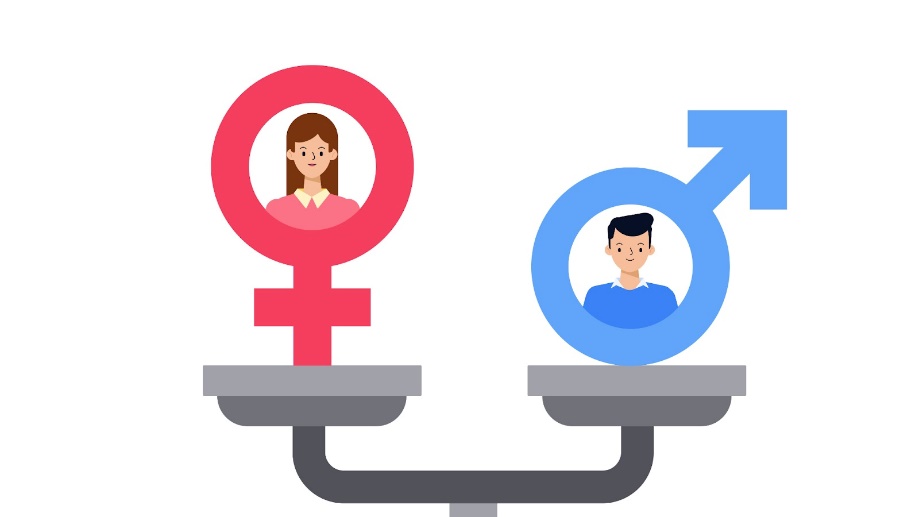
Data ordinal adalah jenis pengelompokan data yang memiliki urutan atau harus disusun secara berurutan dengan mekanisme peringkat. Perhatikan data rating yang ada pada gambar berikut.



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 mana yang memiliki rating tertinggi. Nah informasi tersebut merupakan contoh dari data ordinal bahwa jenis data ini dapat diurutkan berdasarkan peringkat.

***Data Nominal***

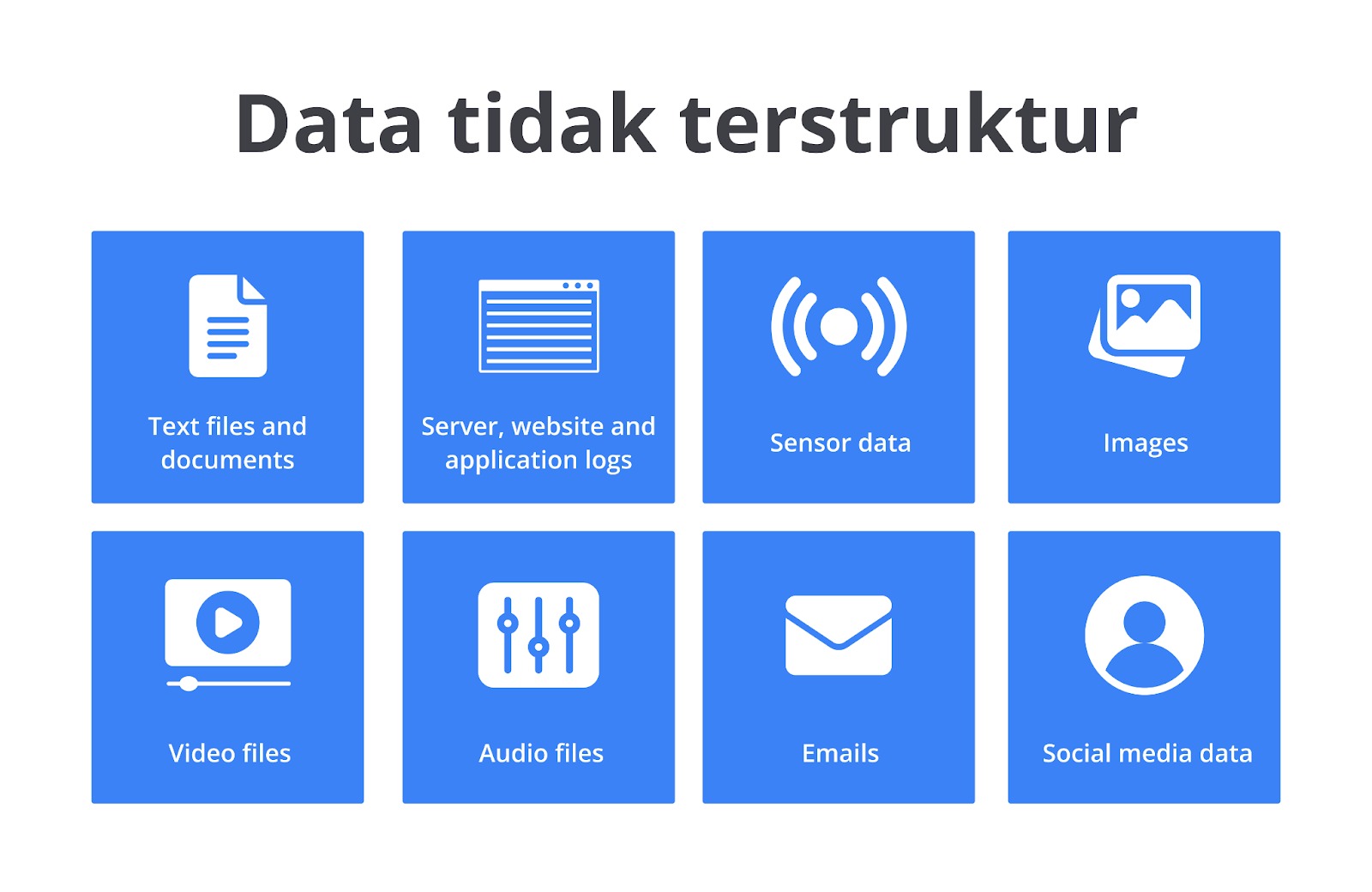
Berkebalikan dengan data ordinal, data nominal adalah jenis pengelompokan data yang tidak memiliki keterkaitan dengan data lainnya dan tidak memiliki arti khusus. Jadi, data ini dapat dibedakan tanpa harus mengurutkan atau dibandingkan dengan data lainnya.



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.



**Kriteria data untuk AI**

Setelah Anda mendalami data untuk AI sebelumnya, sampailah kita pada bagian yang paling penting untuk membangun sebuah sistem AI berdasarkan data yang kita miliki. Seperti pada judulnya, kali ini kita akan membahas 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)**

Pada materi sebelumnya, Anda tentu sudah membaca bahwa data sangatlah penting untuk membangun AI. Nah, sekarang mari kita usut lebih dalam mengapa data memiliki peran yang penting dalam pembangunan AI melalui 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.



Dari gambar di atas kita bisa mendapatkan kesimpulan bahwa keluaran dari sistem AI yang kita bangun sangat bergantung pada data masukkan yang diterima. Jika kita memiliki data masukan 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 yang menyebabkan data memiliki kualitas yang buruk.

**Masalah dalam Data**

Pada tahap ini, kita akan membahas permasalahan yang ada supaya dapat memenuhi kriteria data untuk AI. Lalu, apa saja permasalahan yang sering terjadi pada data dan bagaimana cara menanganinya? Mari kita pelajari beberapa permasalahan yang umum terjadi pada 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.

Sampai di sini, Anda hanya perlu tahu bahwa data yang bermasalah dapat diperbaiki sehingga menghasilkan data yang lebih baik. Untuk menangani permasalah pada data, Anda akan mempelajarinya secara lebih detail di modul berikutnya pada data *preprocessing*. Tetap semangat!

**Infrastruktur Data di Industri**

Data infrastruktur mengacu kepada *hardware*, *software*, dan teknologi jaringan yang digunakan untuk mendukung proses pengelolaan data. Mengapa kita harus mengenal data infrastruktur? Setelah kita mengenal data, alangkah baiknya jika kita mempelajari juga pengelolaan data, penyimpanan, hingga proses analisis data.

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 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.

**[Story] Apa yang Diperlukan untuk Membuat AI?**

Di kisah sebelumnya, Kai masih memiliki pertanyaan yang sangat mengganggu pikirannya. Ia telah mengetahui AI, contoh penerapannya, hingga alur untuk membangun AI. Namun, ia belum mengetahui bahwa AI membutuhkan data agar dapat menjalankan tugasnya dengan baik.

Dengan rasa penasaran serta keinginan untuk mempermudah kehidupannya yang sangat tinggi, Kai membawa permasalahannya ini ke lingkungan universitas. Hingga akhirnya ia bertanya ke salah satu dosennya dengan pertanyaan, “Apa yang membuat AI dapat mengerjakan tugasnya?”.

Dosennya menjawab bahwa AI bekerja dengan menggabungkan data dan melakukan pengolahan sehingga memungkinkan komputer untuk belajar dan melakukan tugasnya secara otomatis berdasarkan pola atau fitur dalam data. Kemudian, dosen tersebut menyarankan Kai untuk mempelajari data terlebih dahulu agar dapat membangun AI dengan baik.

Selepas kuliah, Kai mempelajari hal yang disarankan oleh dosennya tersebut. Ia menemukan fakta di internet yang relevan dengan perkataan dosennya. Fakta tersebut mengatakan bahwa data yang digunakan pada pembangunan AI merupakan hal yang nantinya dapat dipelajari dan dilakukan oleh komputer untuk mengerjakan tugasnya.

Saat ini, Kai sudah mengetahui bahwa ia perlu mempelajari data lebih dalam. Mulai dari pengenalan data hingga tipe-tipe data. Setelah mengetahui ilmu tentang data, motivasinya bertambah berkali-kali lipat. Karena saat ini dia sudah tahu apa yang harus dilakukan untuk memulai membangun sebuah sistem yang dapat membantu tugas yang dimilikinya.

Seiring berjalannya waktu, Kai menemukan sebuah permasalahan bahwa data yang digunakan untuk membangun AI tidak bisa dipakai begitu saja. Kualitas data yang digunakan pada proses pembangunan AI sangat memengaruhi hasil dari sistem yang dibuat.

Berangkat dari permasalahan kualitas data, Kai menemukan bahwa masalah yang muncul pada data sangatlah beragam. Lalu, ia mengetahui beberapa permasalahan yang umum terjadi, seperti data yang tidak relevan (*incorrect*), data berbeda dengan yang lain (*outlier*), data duplikat, data kosong, data yang tidak benar, dan masih banyak lagi.



Akhirnya, Kai mulai merasakan bahwa peran data pada proses pengembangan AI sangatlah penting. Hal itu membuatnya tidak sabar untuk melakukan implementasi berdasarkan data toko yang ia miliki agar dapat membantu meringankan tugasnya.

Penasaran apa yang akan dibangun Kai pada kisah selanjutnya? Simak terus perjalanan Kai hingga ia dapat melakukan semua tugas di tokonya secara otomatis ya!