**Pengantar Pertimbangan dalam Pengolahan Data**



Ini merupakan salah satu ungkapan dari *uncle* Ben dalam seri *Spiderman*. Ungkapan tersebut mengingatkan kita bahwa semakin besar kuasa yang kita miliki maka semakin besar pula tanggung jawabnya. Hal ini haruslah menjadi pengingat bagi Anda para calon praktisi data profesional ketika mengambil keputusan dalam proses analisis data.

Seperti yang telah kita bahas sebelumnya, banyak hal yang bisa dilakukan dengan data, mulai dari mengidentifikasi peluang, mengoptimalkan proses yang ada, hingga membantu kita dalam membuat keputusan bisnis yang tepat. Oleh karena itu, tidak berlebihan jika menyebut data sebagai salah satu sumber daya penting dan *powerful tool* yang ada saat ini.

Sebagai seorang yang mengolah data, kita memiliki tanggung jawab yang sangat besar. Oleh karena itu, banyak ketentuan dan pertimbangan yang harus kita ikuti dalam proses pengolahan data di industri. Nah, pada materi kali ini, kita akan fokus membahas berbagai ketentuan dan pertimbangan dalam proses pengolahan data.

*First of all*, mari kita melihat berbagai masalah dan tantangan yang kita hadapi dalam proses pengolahan data di industri.

**Tantangan dalam Pengolahan Data**

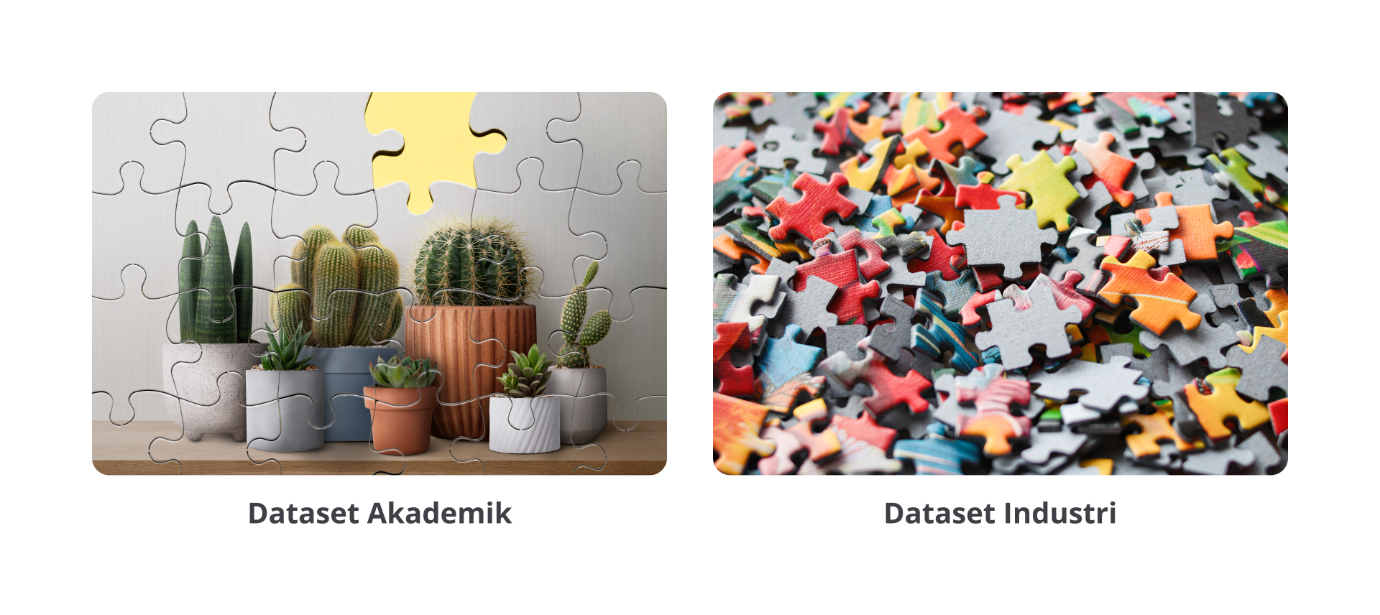
Sejauh ini, Anda telah memahami berbagai konsep dasar dalam analisis data serta *descriptive statistics*. Namun, pernahkah Anda bertanya seperti apa data yang umum dijumpai di industri?

Mungkin sebagian dari Anda pernah mendengar tentang beberapa dataset akademik yang umum digunakan sebagai bahan latihan para calon praktisi data masa depan, seperti *iris dataset*, *titanic dataset*, dll. Seluruh dataset tersebut memang dikhususkan sebagai bahan latihan dan *benchmark* dalam penelitian sehingga bentuknya sangat sederhana. Selain itu, dataset seperti ini umumnya sudah dalam keadaan bersih dan siap untuk digunakan dalam proses analisis data ataupun proses lanjutan lainnya.

|  |
| --- |
| Dataset merupakan kumpulan data yang dikhususkan untuk menyelesaikan tugas tertentu. Dalam kelas ini, kita menggunakan istilah data dan dataset secara bergantian (*interchangeably*). |

Mungkin beberapa dari Anda membayangkan data yang dijumpai di industri memiliki bentuk yang mirip seperti dataset akademik. Anggapan seperti itu tentunya kurang tepat karena data yang kita jumpai di industri sangatlah kompleks dan umumnya masih sangat kotor sehingga perlu dilakukan tahap pembersihan data (data cleaning) terlebih dahulu sebelum digunakan. *FYI*, lebih dari 50% waktu seorang praktisi data profesional dihabiskan untuk membersihkan data [6].

Selain itu, pada proses analisis data, kita sering kali harus menggunakan lebih dari satu dataset dengan ukuran data yang sangat besar untuk menjawab sebuah masalah bisnis. Jika diibaratkan sebagai sebuah puzzle, dataset di industri bisa digambarkan sebagai puzzle yang sangat kompleks dan sulit dicari polanya, sedangkan dataset akademik dapat diwakilkan sebagai puzzle sederhana yang mudah untuk dicari polanya.



Apakah data menjadi satu-satunya tantangan yang akan Anda hadapi di industri? Tentunya tidak, ada banyak tantangan lain yang akan Anda jumpai di industri salah satunya adalah kompleksitas dari kebutuhan bisnis. Kebutuhan bisnis di industri umumnya dipengaruhi banyak hal, mulai dari keadaan pasar, keberadaan kompetitor, hingga keadaan makro ekonomi dan tentunya masih banyak lagi faktor yang lain. Semua faktor tersebut membuat kebutuhan bisnis di industri menjadi sangat dinamis. Inilah tantangan lain yang harus dihadapi oleh seorang praktisi data profesional.

Selain tantangan terkait kebutuhan bisnis dan data, seorang praktisi data juga dihadapkan dengan banyak tantangan lain seperti privasi dan etika dalam pengolahan data serta keamanan dari data itu sendiri. Saat ini, ada banyak regulasi yang mengatur kedua hal tersebut dan sebagai calon praktisi data yang Andal, kita harus mematuhi semua regulasi tersebut.

Nah, itulah gambaran umum mengenai tantangan yang akan Anda hadapi sebagai calon praktisi data profesional. Tentunya tantangan tersebut bukanlah hal yang mudah untuk dihadapi. Namun, jangan khawatir. Latihan yang ada pada kelas ini telah didesain untuk membiasakan Anda bekerja menghadapi kasus yang semirip mungkin dengan yang ada di industri. Jadi, tetap semangat, ya!

**Bias dalam Data**

Seperti yang telah kita bahas pada beberapa materi sebelumnya, tujuan utama dari proses analisis data ialah mencari *insight* dan *meaningful story* dari suatu data. Sebuah cerita yang baik selalu hadir dengan banyak tantangan dan konflik. Data juga seperti itu, ia selalu membawa banyak tantangan. Tugas kita sebagai pemeran utama ialah menjawab tantangan tersebut. Nah, materi kali ini akan mengajak Anda mengenal salah satu tantangan atau masalah yang umum dijumpai dalam sebuah data yaitu bias.

*By definition*, bias merupakan sebuah prasangka dan kecenderungan yang mendukung atau menentang suatu hal, individu, atau kelompok lain dengan cara yang kurang adil. Ia bisa terjadi secara sengaja maupun tidak sengaja. Bias merupakan salah satu masalah yang sering dijumpai di kehidupan sehari-hari termasuk dalam sebuah data.

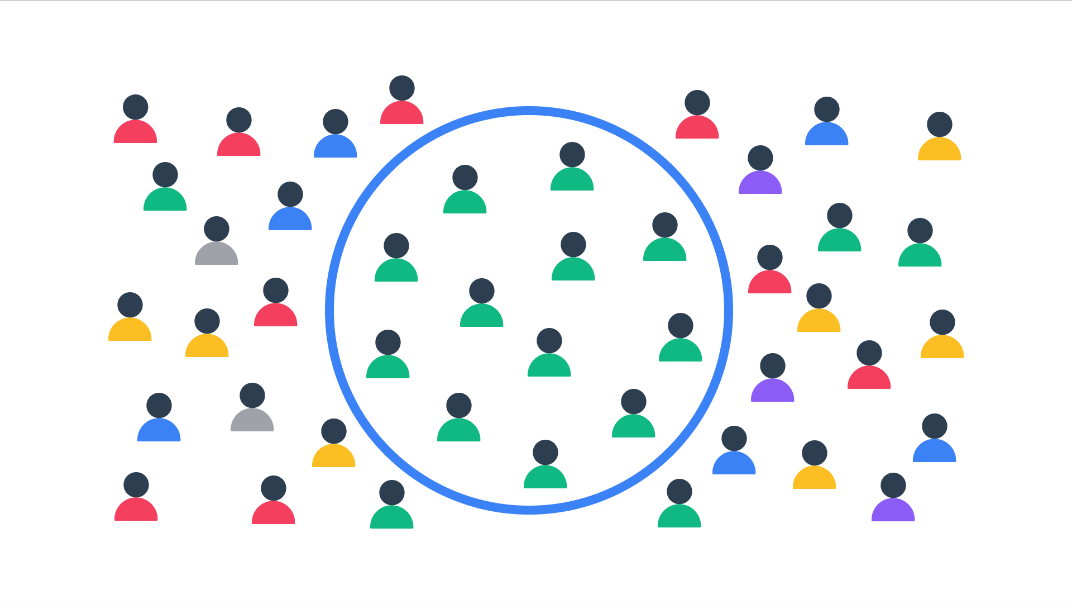
Bias dalam data umumnya didefinisikan sebagai eror atau galat yang secara sistematis mengarahkan hasil menuju arah tertentu. Bias dalam data umumnya muncul disebabkan banyak faktor salah satunya adalah karena pengambilan sampel yang kurang merepresentasikan populasi secara keseluruhan [7].

Sebagai seorang calon praktisi data yang andal, kita perlu mempertimbangkan faktor bias dalam keseluruhan proses analisis data mulai dari tahap pengumpulan hingga mengkomunikasikan hasil analisis data. Hal ini sangat penting untuk mencegah hasil analisis data yang tidak adil yang mungkin berdampak negatif terhadap perkembangan sebuah bisnis. Oleh karena itu, sangat penting bagi kita untuk mengetahui cara mengatasi atau menghindari bias dalam sebuah data.

**Cara Mengatasi Bias dalam Data**

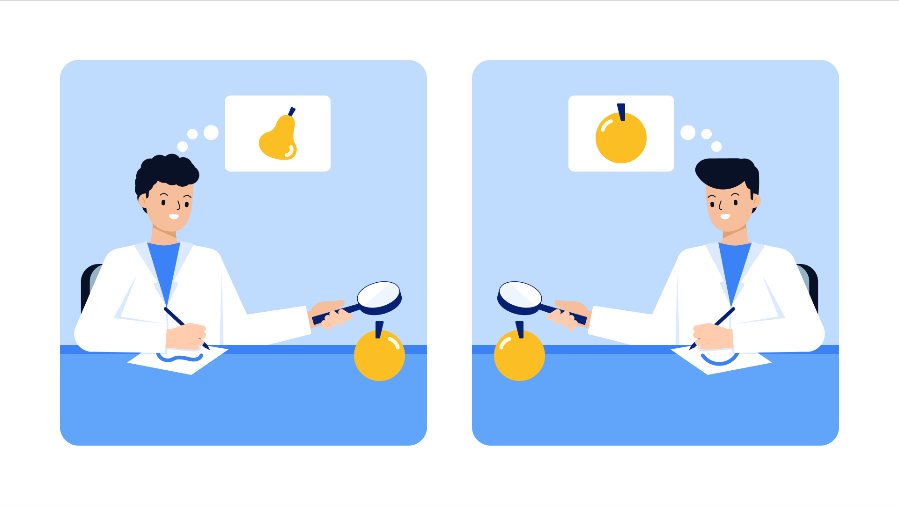
Oke, pada tahap ini, kita telah paham betapa berbahayanya bias dalam sebuah data. Sekarang kita akan membahas berbagai teknik dalam mengatasinya. Teknik dalam mengatasi bias akan berbeda-beda tergantung tipe dari bias tersebut. Nah, secara umum bias dalam data dapat dikategorikan dalam empat tipe, yaitu *sampling bias*, *observer bias*,*interpretation bias*, dan *confirmation bias*[7].

* **Sampling bias**  
  Sampling bias terjadi ketika sampel tidak mewakili populasi secara keseluruhan. Ini merupakan salah satu tipe bias yang paling sering terjadi. Bias jenis ini bisa muncul karena data sampel yang kita miliki kurang seluruh kelompok data yang ada atau hanya merepresentasikan kelompok data tertentu.



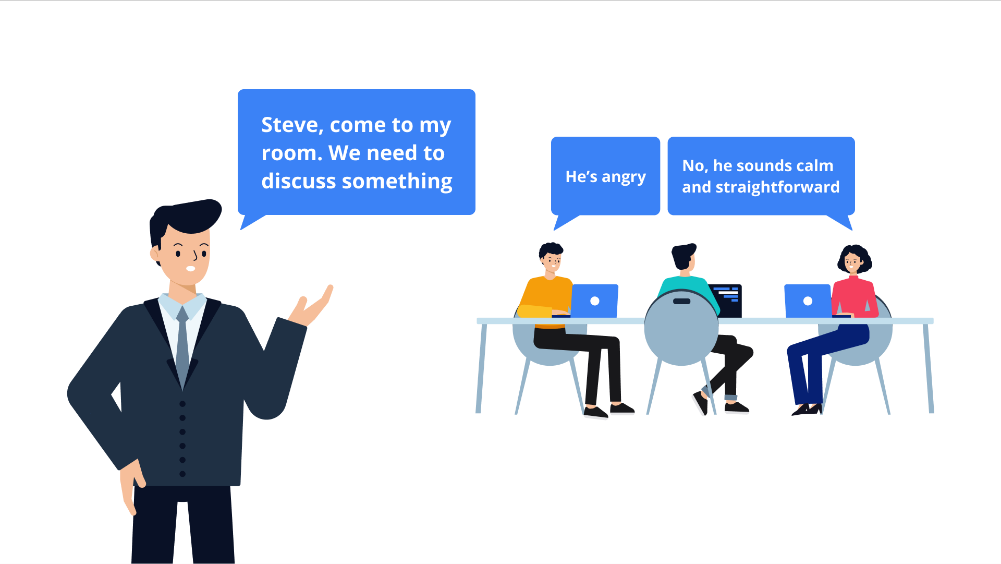
Bias seperti ini dapat dihindari menggunakan teknik *random sampling* ketika mengumpulkan sebuah data. Teknik ini memungkinkan semua bagian dari populasi mendapat kesempatan yang sama untuk terpilih.

* **Observer bias**  
  Tipe bias selanjutnya ialah observer bias atau sering juga disebut sebagai *experimenter bias*. Bias tipe ini muncul karena adanya kecenderungan yang berbeda dari setiap individu ketika melakukan observasi.



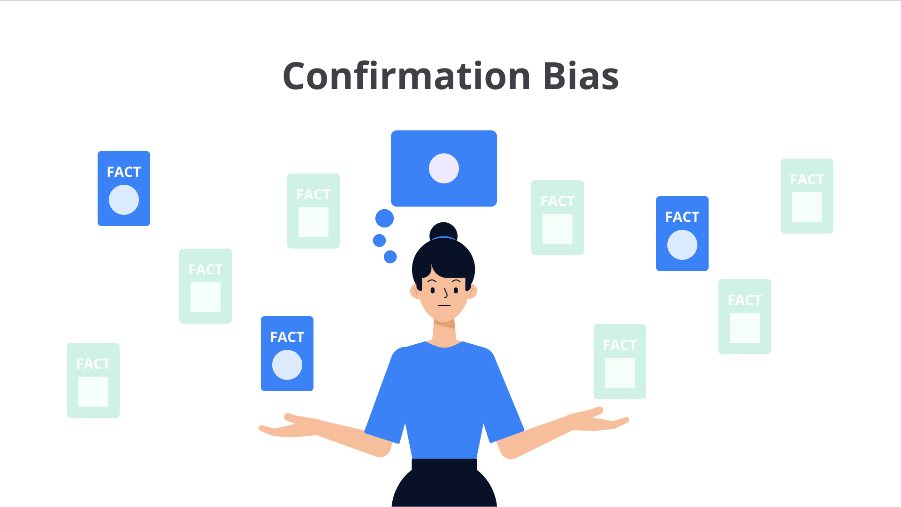
Bias jenis ini dapat dihindari dengan melakukan observasi berulang. Sebagai contoh, untuk menghindari observasi bias ketika melakukan pengukuran, kita bisa melakukan proses pengukuran secara berulang dan mengambil nilai rata-ratanya sebagai hasil akhir.

* **Interpretation bias**  
  Interpretation bias merupakan bias yang terjadi karena adanya kecenderungan kita dalam menginterpretasikan situasi ambigu hanya ke dalam dua keadaan (hitam dan putih). Hal ini berpotensi melahirkan bias karena setiap individu memiliki kecenderungan yang berbeda dalam menginterpretasi situasi yang ambigu.



Hal ini dapat dihindari dengan cara melibatkan orang lain untuk mengonfirmasi hasil interpretasi yang kita buat. Selain itu, kita juga dapat menambahkan parameter atau nilai lain yang merepresentasikan keadaan ambigu.

* **Confirmation bias**  
  Tipe bias terakhir adalah confirmation bias. Bias ini muncul karena adanya kecenderungan kita dalam mencari atau menafsirkan informasi untuk mengonfirmasi keyakinan yang sudah ada sebelumnya. Tipe bias ini akan sangat berbahaya karena dapat menjebak kita dalam *echo chamber*(keadaan yang membuat seseorang hanya menerima informasi dan opini yang sesuai dengan yang mereka percayai).

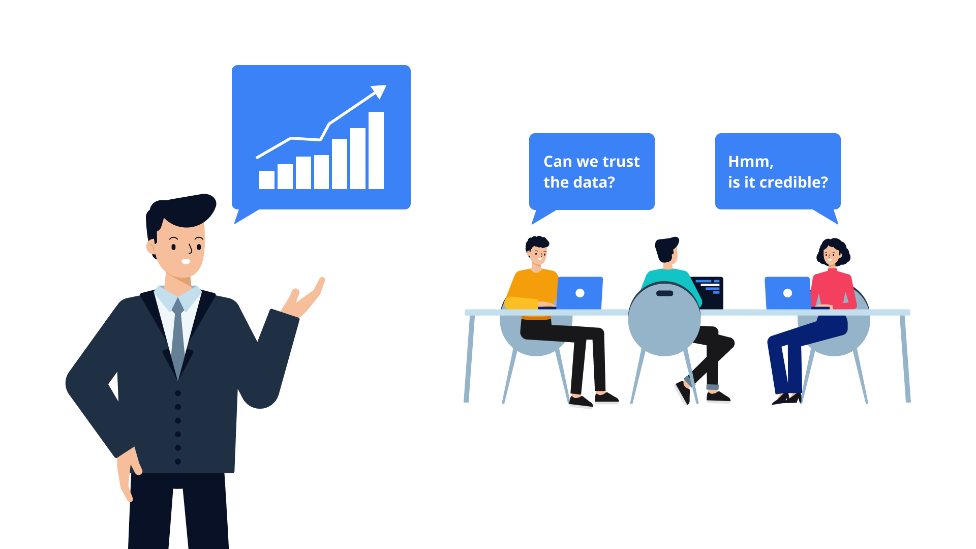


Untuk mengatasi bias seperti ini, kita perlu selalu menempatkan diri senetral mungkin dan mencari fakta lain yang tidak hanya mengonfirmasi sebuah pernyataan (atau ide) melainkan juga menyanggah pernyataan tersebut. Oleh karena itu, sebagai calon praktisi data yang andal, kita harus selalu menempatkan diri senetral mungkin dalam setiap proses analisis data untuk menghindari *confirmation bias*.

**Data Credibility**

Pada materi sebelumnya, kita telah membahas tentang bias dalam data dan impaknya terhadap hasil analisis. Hal ini tentunya menjadi pengingat bagi kita tentang pentingnya untuk memastikan data yang digunakan dalam proses analisis merupakan data yang berkualitas dan dapat dipercaya (kredibel).

“*Hm…*, bagaimana cara menilai dan mengidentifikasi data yang berkualitas serta kredibel?” Jika ini yang terlintas, selamat Anda berada di tempat yang tepat. Pada materi kali ini, kita akan membahas segala hal yang berkaitan dengan kualitas dan kredibilitas dari sebuah data.



*First of all*, kita perlu memahami definisi kredibilitas data. Ia merupakan bagian dari *data quality assessment* yang digunakan sebagai sebuah tolok ukur untuk menilai tingkat kepercayaan terhadap suatu data.

Tingkat kepercayaan dari sebuah data akan sangat berdampak terhadap kualitas dan keakuratan hasil analisis yang akan dibuat. Tingkat kepercayaan ini tentunya dipengaruhi oleh banyak faktor, seperti sumber data, kelengkapan data, keakuratan data, dll.

Nah, untuk memastikan kredibilitas data, kita perlu melihat keseluruhan faktor tersebut. Berikut merupakan beberapa poin sebagai patokan untuk memastikan data yang digunakan berkualitas dan dapat dipercaya [7].

* **Reliabilitas data**  
  Reliabilitas merupakan salah satu faktor penting untuk menilai kualitas dan kredibilitas dari sebuah data. Data disebut reliabel jika memuat informasi yang akurat, lengkap, dan tidak bias.  
    
  Bisa Anda bayangkan betapa berbahayanya jika proses analisis dilakukan pada data yang tidak reliable. Proses tersebut akan menghasilkan analisis yang tidak dapat dipercaya dan tentunya akan berdampak terhadap perkembangan sebuah bisnis. Oleh karena itu, kita sangat disarankan untuk selalu menilai reliabilitas dari sebuah data sebelum menjalankan proses analisis.
* **Orisinalitas data**  
  Selain reliabilitas, orisinalitas dari sebuah data juga merupakan komponen kunci dalam menilai kualitas dan kredibilitas dari sebuah data. Hal ini erat kaitannya dengan *ownership* (kepemilikan) dan sumber data.  
    
  Ketika bekerja menggunakan data dari eksternal, kita harus selalu menilai keaslian data tersebut dengan membandingkan dengan sumber aslinya. Selain itu, kita juga bisa memastikan reputasi dari sumber data tersebut dan melihat apakah ia bagian dari organisasi yang kredibel?
* **Komprehensif**  
  Poin selanjutnya yang tidak kalah penting ialah komprehensif. Poin ini sangat penting untuk memastikan data yang digunakan memiliki semua informasi yang dibutuhkan untuk menjawab pertanyaan dan tantangan bisnis.  
    
  Sebuah data dikatakan berkualitas jika memuat semua informasi yang dibutuhkan untuk melakukan analisis data. Kelengkapan informasi ini tentunya akan sangat berpengaruh terhadap keakuratan dari suatu hasil analisis.
* **Aktualitas data**  
  *Last*,*but not least* adalah aktualitas data. Seperti yang telah kita bahas sebelumnya, data di industri umumnya bersifat dinamis atau terus berubah. Jika kita menggunakan data lama, informasi di dalamnya mungkin akan *outdated* sehingga berdampak terhadap hasil analisis yang dibuat. Oleh karena itu, sangat penting bagi kita untuk memastikan proses analisis dilakukan menggunakan data terbaru.

Itulah keempat poin yang harus kita perhatikan untuk menilai data yang berkualitas dan dapat dipercaya (kredibel). Semoga pengetahuan ini dapat menjadi bekal bagi Anda dalam menilai sebuah data nantinya.

**Privasi dan Etika dalam Pengolahan Data**

Pada tahap ini, Anda telah mengenal beberapa masalah dan tantangan yang umum dijumpai dalam proses pengolahan data di industri, yaitu bias dan kredibilitas data. Hal ini tentunya belum mencakup semua tantangan yang akan Anda hadapi di industri.

Nah, pada materi kali ini, kita akan membahas tentang privasi dan etika dalam proses pengolahan data. Pada beberapa dekade terakhir, kedua topik tersebut telah menjadi salah satu pembahasan hangat dalam dunia data. Hal ini merupakan imbas dari maraknya pemanfaatan data di industri sehingga mulai banyak yang menyadari potensi bahaya dalam praktik tersebut seperti pelanggaran privasi.



Untuk mencegah bahaya tersebut, mulai muncul banyak studi terkait *data ethics*. Ia merupakan kumpulan acuan standar dalam menilai benar dan salah sebuah proses pengolahan data. Berikut merupakan beberapa aspek penting dalam data ethics [7].

* **Ownership**  
  Aspek ini menyatakan bahwa kepemilikan data berada pada setiap individu yang menghasilkan data tersebut. Sebagai pemilik data, individu memiliki kontrol penuh akan seluruh proses dan penggunaan data tersebut.
* **Transparency**  
  Seluruh tahapan dan algoritma dalam proses pengolahan data harus transparan dan dipahami oleh setiap individu yang menghasilkan data tersebut.
* **Consent**  
  Setiap individu harus mengetahui secara eksplisit terkait penggunaan dan alasan suatu data dikumpulkan. Hal ini dilakukan sebelum individu tersebut menyetujui pengumpulan data tersebut.
* **Currency**  
  Jika data digunakan untuk transaksi keuangan, setiap individu yang menghasilkan data harus mengetahui dengan jelas detail terkait transaksi tersebut.
* **Privacy & Openness**  
  Organisasi yang mengumpulkan data bertanggung jawab dalam menjaga setiap informasi yang bersifat pribadi pada setiap proses pengolahan data. Selain itu, organisasi tersebut harus menyediakan akses terhadap penggunaan dan pemanfaatan data (non private) secara bebas.

Kelima aspek tersebut banyak digunakan sebagai dasar dalam pembuatan regulasi dalam pengolahan data. Yuk, simak penjelasan berikut untuk memahami lebih berbagai regulasi dalam pengolahan data.

**Regulasi dalam Pengolahan Data**

Seperti yang telah kita bahas sebelumnya, privasi merupakan salah satu isu penting yang harus kita pertimbangkan dalam menjaga hak setiap individu. Oleh karena itu, banyak lembaga atau instansi pemerintah di seluruh dunia mulai memperhatikan isu tersebut dan membuat regulasi untuk mengaturnya. Berikut merupakan beberapa regulasi yang dibuat untuk menjaga data privasi milik pengguna dan memastikan penerapan data ethics.

* **General Data Protection Regulation (GDPR)**  
  Salah satu regulasi terkait privasi pengguna yang sering digunakan di industri adalah General Data Protection Regulation (GDPR). Ia merupakan regulasi yang dibuat oleh Uni Eropa untuk mengatur dan melindungi pemanfaatan data pengguna dalam lingkup European Economic Area (EEA). Regulasi ini memberikan kontrol terhadap individu atas penggunaan data pribadi mereka dan mewajibkan perusahaan untuk melindungi data tersebut. Selain itu, regulasi ini hanya mengizinkan pemanfaatan data berdasarkan persetujuan pengguna. Aturan lengkap terkait regulasi ini dapat Anda lihat pada tautan berikut: [GDPR](https://gdpr.eu/tag/gdpr/).
* **Peraturan Menteri (Permen) No 20 Tahun 2016 tentang Perlindungan Data Pribadi (PDP)**

Sebagai negara yang berdaulat, Indonesia juga memiliki sebuah regulasi yang mengatur dan melindungi data pengguna. Regulasi ini diatur dalam bentuk Permen No 20 Tahun 2016 tentang Perlindungan Data Pribadi (PDP). Peraturan ini mewajibkan setiap penyelenggara sistem elektronik untuk mempunyai regulasi internal terkait perlindungan data pribadi pengguna dan memungkinkan pemilik data untuk mengubah dan memperbarui data tersebut. Selain itu, Permen ini juga hanya mengizinkan penggunaan data sesuai persetujuan pengguna atau pemilik data. Anda dapat membaca PerMen (Peraturan Menteri) ini secara lengkap melalui tautan berikut: Perlindungan Data Pribadi Dalam Sistem Elektronik.

Nah, itulah beberapa regulasi yang harus selalu kita pertimbangkan dalam setiap tahap pengolahan dan analisis data. Pada materi berikutnya kita akan membahas tentang salah satu isu yang tidak kalah penting yaitu data security atau keamanan data.

**Keamanan Data**

Pada materi sebelumnya, kita telah membahas tentang privasi dan etika dalam pengolahan data. Selain itu, kita juga telah membahas berbagai regulasi yang mengatur penggunaan data dan memastikan data tersebut memenuhi seluruh aspek dalam *data ethics* terutama terkait privasi pengguna.

Ketika membahas tentang data ethics dan privasi pengguna, terdapat satu aspek penting yang dibutuhkan untuk memastikan kedua hal tersebut terpenuhi. Apakah Anda bisa menebaknya? Betul sekali, aspek tersebut ialah *data security* atau keamanan data.

Data security dapat didefinisikan sebagai upaya dalam menjaga dan melindungi informasi digital (data) dari berbagai pihak yang tidak bertanggung jawab dalam seluruh proses pengolahannya. Hal ini mencakup keamanan fisik *hardware* (*storage device*) serta keamanan dari *software* atau aplikasi yang bersangkutan. Selain itu, ia juga mencakup seluruh kebijakan dan prosedur dalam sebuah organisasi [8].

Seperti yang kita ketahui, proses analisis data umumnya membutuhkan data yang cukup banyak. Hal inilah yang menjadi tantangan tersendiri bagi kita para praktisi data untuk menyeimbangkan kebutuhan analisis dengan tetap menjaga privasi pengguna.

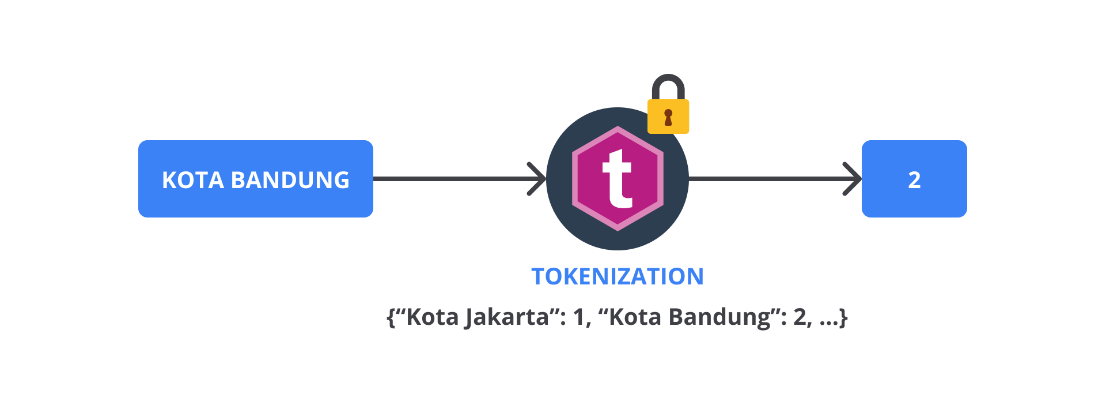
Beruntungnya terdapat beberapa teknik data security yang dapat kita terapkan guna menyeimbangkan kedua kebutuhan tersebut. Pada kesempatan ini kita akan membahas dua diantaranya yaitu seperti berikut.

* **Encryption**  
  Encryption merupakan salah satu metode yang paling sering digunakan dalam lingkup data security. Ia merupakan teknik yang memanfaatkan algoritma tertentu untuk mengubah data menjadi bentuk *unusable* bagi individu atau aplikasi yang tidak memiliki akses terhadap algoritma tersebut.



Pada penerapnnya, kita bisa menganggap algoritma tersebut sebagai *key* (kunci) untuk membalikkan (*reverse*) proses encryption. Oleh karena itu, jika memiliki akses terhadap algoritma tersebut, kita tetap bisa menggunakan data tersebut dalam bentuk originalnya.

* **Tokenization**  
  Metode lain yang bisa digunakan ialah tokenisasi. Metode ini memungkinkan kita untuk mengubah elemen tertentu dalam sebuah data menjadi sebuah data random yang berperan sebagai token.



Pada metode ini, data asli beserta *token mapping* akan disimpan dalam lokasi yang berbeda. Hal ini memungkinkan kita membatasi akses terhadap data asli dan memastikan keamanannya dengan menyimpan data tersebut di dalam lokasi yang berbeda.

Oke, itulah dua metode yang dapat diterapkan untuk memastikan keamanan data dan menjaga privasi pengguna. Tentunya masih ada beberapa metode lain untuk melakukannya, tetapi untuk saat ini Anda cukup memahami dua metode dasar tersebut terlebih dahulu.