**Mempersiapkan Data untuk Eksplorasi**

**Data yang tidak bias dan objektif**

Secara biologis otak kita dirancang untuk berpikir secara efisien dan membuat penilaian dengan cepat. Bias telah berkembang menjadi suatu preferensi terkait mendukung atau menentang seseorang, sekelompok orang, atau suatu hal. Hal ini dapat terjadi secara sadar atau secara tidak sadar.

Bias data adalah jenis kesalahan yang secara sistematis mengarahkan hasil menuju arah tertentu. Mungkin pertanyaan yang ada dalam survei memiliki kecenderungan tertentu untuk memengaruhi jawaban, atau mungkin kelompok sampelnya tidak mewakili populasi yang sedang diteliti.

**Memahami Bias Dalam Data**

* **Bias Sampling**, adalah ketika sampel tidak mewakili populasi secara keseluruhan. Misalnya, jika Anda melakukan penelitian mengenai commuter, dan melakukan survei hanya kepada orang-orang yang berjalan di trotoar, anda akan kehilangan input dari orang-orang yang mengendarai sepeda, naik kendaraan, atau naik MRT.
* **Bias observer**, yang terkadang disebut sebagai bias eksperimen (experimenter bias) atau bias penelitian. Pada dasarnya, ini adalah kecenderungan di mana orang yang berbeda, mengamati hal yang berbeda. Seperti para ilmuwan menggunakan banyak pengamatan dalam pekerjaan mereka, seperti ketika mereka melihat bakteri melalui mikroskop untuk mengumpulkan data. Saat dua ilmuwan melihat ke dalam mikroskop yang sama namun melihat hal yang berbeda.
* **Bias interpretasi**. Kecenderungan untuk selalu menafsirkan situasi yang ambigu dengan cara yang positif, atau negatif. Berikut adalah contohnya. Katakanlah Anda sedang makan siang dengan seorang teman dan Anda mendapatkan pesan suara dari atasanmu, yang memintamu untuk melakukan panggilan telepon. Anda meletakkan telepon dengan gusar, Anda yakin bahwa dia marah, dan Anda berada dalam masalah. Tetapi ketika Anda memutarkan pesan suara itu untuk didengarkan oleh teman Anda, teman Anda tidak mendengar adanya kemarahan, teman Anda merasa bos Anda terdengar tenang dan lugas. Bias interpretasi, dapat menyebabkan dua orang yang melihat atau mendengar hal yang sama, menafsirkannya dengan berbagai cara yang berbeda, karena mereka memiliki perbedaan latar belakang, dan pengalaman.
* **Bias konfirmasi,** adalah kecenderungan untuk mencari, atau menafsirkan informasi dengan cara mengkonfirmasi keyakinan yang sudah ada sebelumnya. Seseorang mungkin ingin mengkonfirmasi firasat, sehingga dia hanya memperhatikan hal-hal yang mendukungnya, mengabaikan semua petunjuk lain. Hal ini terjadi dalam kehidupan sehari-hari. Kita bisa mendapatkan berita dari situs web tertentu karena para penulisnya memiliki keyakinan yang sama dengan kita, atau kita bersosialisasi dengan orang-orang karena kita tahu bahwa mereka memiliki pandangan yang sama.

**Sumber Data Yang Baik, R-O-C-C-C (Reliable, Original, Comperhensive, Current and Cited)**

* **Reliable (**dapat diandalkan), sumber data yang baik harus dapat diandalkan. Dengan data yang dapat Anda percaya, Anda mendapatkan informasi akurat, lengkap dan tidak bias yang telah diperiksa dan terbukti layak untuk digunakan.
* **Original** (asli), ada kemungkinan besar Anda akan menemukan data melalui sumber pihak kedua atau pihak ketiga. Untuk memastikan Anda menggunakan data yang baik, pastikan untuk melakukan validasi pada sumber aslinya
* **Comperhensive (**komprehensif),Sumber data terbaik berisi semua informasi kritis yang dibutuhkan untuk menjawab pertanyaan atau menemukan solusinya.
* **Current** (saat ini), kegunaan data akan berkurang seiring berjalannya waktu. Jika Anda mengundang semua klien saat ini untuk acara bisnis, Anda tidak akan menggunakan daftar klien yang telah berusia 10 tahun. Hal yang sama akan berlaku untuk data. Sumber data yang terbaik adalah yang terkini dan relevan dengan tugas yang ada
* **Cited** (dikutip). Jika Anda pernah menceritakan kepada teman Anda bahwa Anda mendengar kelanjutan untuk film tertentu sedang dikerjakan, Anda telah mengambil kutipan dari sebuah sumber. Kutipan membuat informasi yang Anda berikan menjadi lebih kredibel.

**Sumber Data Yang Tidak Baik (Not-ROCCC)**

* **Not Reliable** (tidak dapat diandalkan). Data yang buruk tidak dapat dipercaya karena data tersebut tidak akurat, tidak lengkap, atau bias. Ini bisa jadi data yang memiliki bias pemilihan sampel karena tidak mencerminkan populasi secara keseluruhan. Atau bisa juga visualisasi data dan grafik yang menyesatkan.
* **Not Original** (tidak asli). Jika Anda tidak dapat menemukan sumber data yang asli dan Anda hanya mengandalkan informasi dari pihak kedua atau pihak ketiga, itu adalah peringatan bahwa Anda mungkin perlu sangat hati-hati dalam memahami data Anda.
* **Not Comprehensive (**tidak komprehensif). Sumber data yang buruk adalah yang tidak memiliki informasi penting yang diperlukan untuk menjawab pertanyaan atau menemukan solusinya. Yang lebih buruk, mereka mungkin memiliki elemen human error.
* **Not Current (**tidak terkini). Sumber data yang buruk adalah sumber data yang sudah ketinggalan zaman dan tidak relevan. Banyak sumber data yang terpercaya memperbarui data mereka secara teratur, untuk memastikan bahwa Anda mendapatkan informasi yang terbaru/terkini.
* **Not Cited** (tidak dikutip). Jika sumber yang Anda gunakan tidak pernah dikutip atau diperiksa sebelumnya, maka jangan gunakan sumber tersebut.

**Etika Data**

Satu pandangan praktis adalah bahwa etika mengacu pada standar mengenai benar dan salah yang menentukan apa yang harus dilakukan manusia, biasanya dalam hal hak, kewajiban, manfaat bagi masyarakat, keadilan atau kebajikan tertentu. Sama seperti manusia, data juga memiliki standar hidup. Etika data mengacu pada standar benar dan salah yang menentukan bagaimana data itu dikumpulkan, dibagikan dan digunakan.

Ada banyak aspek yang berbeda dari etika data. Kita akan membahas enam aspek: kepemilikan, transparansi transaksi, persetujuan, nilai uang, privasi, dan keterbukaan.

* **Ownership (Kepemilikan).** Ini akan menjawab pertanyaan mengenai siapa yang memiliki data? Bukan organisasi yang menginvestasikan waktu dan uang yang mengumpulkan, menyimpan, memproses, dan menganalisisnya. Individu-individu lah yang memiliki data mentah yang mereka berikan, dan mereka memiliki kendali utama atas penggunaannya, bagaimana data tersebut diproses dan bagaimana data tersebut dibagikan.
* **Tarnsaction Transparency (Transparansi transaksi)**, yang merupakan gagasan bahwa semua aktivitas pemrosesan data dan algoritma harus dapat dijelaskan sepenuhnya dan dipahami oleh individu yang memberikan datanya.
* **Consent (Persetujuan).** Yaitu hak individu untuk mengetahui detail yang jelas tentang bagaimana dan mengapa data mereka akan digunakan sebelum setuju untuk memberikannya. Mereka harus tahu jawaban atas pertanyaan seperti mengapa data dikumpulkan? Bagaimana data itu akan digunakan? Berapa lama data akan disimpan? Cara terbaik untuk memberikan persetujuan mungkin adalah melalui percakapan antara orang yang memberikan data dan orang yang memintanya.
* **Nilai uang (currency**). Individu harus menyadari bahwa transaksi keuangan yang dihasilkan dari penggunaan data pribadi mereka dan dari skala transaksi tersebut. Jika data Anda membantu mendanai usaha dari perusahaan, Anda harus mengetahui semua hal mengenai usaha tersebut dan diberi kesempatan untuk memilih keluar.
* **Privacy (Privasi)** berarti menjaga informasi dari subjek data dan aktivitasnya saat terjadi transaksi data. Kadang-kadang ini disebut privasi informasi atau perlindungan data. Ini semua tentang akses, penggunaan, dan pengumpulan data. ni berarti seseorang orang seperti Anda atau saya seharusnya memiliki perlindungan dari akses yang tidak sah ke data pribadi kita, kebebasan dari penggunaan yang tidak pantas atas data kita, hak untuk memeriksa, memperbarui, atau memperbaiki data kita, kemampuan untuk memberikan persetujuan untuk menggunakan data kita, dan hak hukum untuk mengakses data kita. Bagi perusahaan, itu berarti menerapkan tindakan-tindakan privasi untuk melindungi data seseorang.
* **Openness (Keterbukaan),** menyediakan akses, penggunaan, dan berbagi data secara bebas. Tapi itu tidak berarti kita mengabaikan aspek lain dari etika data yang kita bahas. Kita harus tetap transparan, menghormati privasi, dan memastikan kita memiliki persetujuan untuk data yang dimiliki oleh orang lain. Ini berarti kita dapat mengakses, menggunakan, dan membagikan data itu jika memenuhi standar yang ketat dan tinggi

**Anonimisasi data**

Informasi identitas pribadi *(Personally identifiable information*), atau PII, adalah informasi yang dapat digunakan secara terpisah atau dengan data lain untuk melacak identitas seseorang. Anonimisasi data adalah proses melindungi data pribadi atau sensitif milik orang dengan menghilangkan informasi semacam itu. Biasanya, anonimisasi data melibatkan pengosongan, hashing, atau menutupi informasi pribadi, seringkali dengan menggunakan kode yang memiliki panjang tetap untuk mewakili kolom data, atau menyembunyikan data dengan nilai yang diubah.

**Peran Anda dalam anonimisasi data**

Organisasi memiliki tanggung jawab untuk melindungi data mereka dan informasi pribadi yang mungkin terkandung dalam data tersebut. Sebagai seorang analis data, Anda mungkin diharapkan untuk memahami data apa yang perlu dianonimkan, namun secara umum Anda tidak akan bertanggung jawab atas anonimisasi data itu sendiri. Pengecualian yang langka mungkin terjadi jika Anda bekerja dengan salinan data untuk tujuan pengujian atau pengembangan. Dalam kasus tersebut, Anda dapat diminta untuk menganonimkan data sebelum Anda bekerja menggunakan data tersebut.

**Jenis data apa yang harus dianonimkan?**

Data kesehatan dan keuangan adalah dua jenis data yang paling sensitif. Industri-industri ini sangat bergantung pada teknik anonimisasi data. Selain itu, taruhannya sangat tinggi. Itu sebabnya data dalam kedua industri ini biasanya melewati proses de-identifikasi, yang merupakan proses yang digunakan untuk menghapus bersih data dari semua informasi identifikasi pribadi.

Anonimisasi data digunakan di hampir setiap industri. Itulah mengapa sangat penting bagi analis data untuk memahami dasar-dasarnya. Berikut adalah daftar data yang sering dianonimkan:

* Nomor telepon
* Nama
* Plat kendaraan dan nomor lisensi kendaraan
* Nomor jaminan sosial
* Alamat IP
* Rekam medis
* Alamat email
* Foto-foto
* Nomor rekening

**Apa itu open data?**

Dalam analitik data, open data adalah bagian dari etika data, yang berkaitan dengan menggunakan data secara etis. Keterbukaan mengacu pada akses, penggunaan, dan berbagi data secara bebas dan gratis. Tapi agar data dianggap terbuka (open), data tersebut harus:

* Tersedia dan dapat diakses oleh publik dalam bentuk dataset yang lengkap
* Disediakan berdasarkan persyaratan yang memungkinkannya untuk digunakan dan didistribusikan kembali
* Mengizinkan partisipasi universal sehingga siapa pun dapat memakai, menggunakan kembali, dan mendistribusikan kembali data tersebut

Data hanya dapat dianggap terbuka ketika memenuhi ketiga standar ini.

**Debat mengenai open data: Data apa yang harus tersedia untuk umum?**

Salah satu manfaat terbesar dari open data adalah database yang kredibel dapat digunakan secara lebih luas. Pada dasarnya, ini berarti bahwa semua data yang baik dapat dimanfaatkan, dibagikan, dan dikombinasikan dengan data lainnya. Hal ini dapat berdampak besar pada kolaborasi ilmiah, kemajuan penelitian, kapasitas analitis, dan pengambilan keputusan. Tetapi penting untuk memikirkan bahwa individu yang diwakili oleh publik, ada open data juga.

**Data pihak ketiga** dikumpulkan oleh entitas yang tidak memiliki hubungan langsung dengan datanya. Anda mungkin ingat mempelajari tentang jenis data ini sebelumnya. Misalnya, pihak ketiga mungkin mengumpulkan informasi tentang pengunjung ke situs web tertentu. Melakukan hal ini memungkinkan pihak ketiga ini untuk membuat profil audiens, yang membantu mereka lebih memahami perilaku pengguna dan menargetkan audiens mereka dengan iklan yang lebih efektif.

**Informasi identitas pribadi (PII)** adalah data yang cukup memungkinkan untuk mengidentifikasi seseorang dan informasi yang diketahui tentang dirinya. Penting halnya untuk menjaga keamanan data ini. PII dapat mencakup alamat seseorang, informasi kartu kredit, nomor jaminan sosial, rekam medis, dan banyak lagi.

Semua orang ingin menyimpan informasi pribadi tentang diri mereka sendiri secara privat. Karena data pihak ketiga sudah tersedia, penting halnya untuk menyeimbangkan keterbukaan data dengan privasi individu.

**Situs dan referensi terkait open data**

Untungnya bagi analis data, ada banyak situs dan referensi tepercaya tentang open data yang tersedia. Penting untuk diingat bahwa bahkan data yang bereputasi baik perlu terus dievaluasi, tetapi situs-situs web ini adalah titik awal yang berguna:

1. [Situs data pemerintah AS: Data.gov](https://data.gov/) adalah salah satu sumber data paling komprehensif di AS. Sumber ini memberi pengguna data dan alat yang mereka butuhkan untuk melakukan penelitian, dan bahkan membantu mereka mengembangkan aplikasi web dan seluler serta mendesain visualisasi data.
2. [U.S. Census Bureau](https://www.census.gov/data.html): Sumber open data ini menawarkan informasi demografis dari pemerintah federal, negara bagian, dan lokal, dan juga entitas komersial di AS.
3. [Open Data Network](https://www.opendatanetwork.com/): Sumber data ini memiliki mesin pencari yang sangat mumpuni dengan filter yang canggih. Di sini, Anda dapat menemukan data tentang topik seperti keuangan, keselamatan publik, infrastruktur, serta perumahan dan pengembangan.
4. [Google Cloud Public Datasets](https://cloud.google.com/): Ada pilihan dataset publik yang tersedia pada Program Dataset Publik Google Cloud (Google Cloud Public Dataset Program) yang dapat Anda temukan sudah dimuat ke BigQuery.
5. [Dataset Search: The Dataset Search](https://datasetsearch.research.google.com/) adalah mesin pencari yang dirancang khusus untuk set data; Anda dapat menggunakan alat ini untuk mencari set data tertentu.