Pada materi sebelumnya Anda telah belajar tentang pengenalan machine learning (ML) dan jenis-jenis library machine learning. Apakah Anda bingung harus melangkah dari manakah untuk memulai proyek ML?

OK Jangan khawatir. Kita akan belajar sedikit demi sedikit ya. Tahap pertama dari proses pengerjaan proyek ML adalah data collecting, yaitu proses pengumpulan data.

“Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”

-Clive Humby, 2006-

Kutipan di atas adalah kalimat terkenal tentang data yang pertama kali disampaikan oleh Clive Humby, seorang matematikawan asal Inggris pada tahun 2006. Kutipan tersebut menjadi sangat populer setelah [The Economist](https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data) mempublikasikan laporan tahun 2017 yang berjudul The World’s most valuable resource is no longer oil, but data.

[](https://www.dicoding.com/academies/184/tutorials/11946?from=10752)

Perangkat cerdas dan internet telah membuat data menjadi berlimpah. Banjir arus data yang terjadi di era digital mengubah sifat persaingan. Perusahaan teknologi raksasa berlomba-lomba mengumpulkan banyak data untuk meningkatkan produknya, menarik lebih banyak pengguna, menghasilkan lebih banyak data, dan seterusnya. Mereka menjangkau seluruh sektor ekonomi: Google bisa melihat apa yang ditelusuri dan dicari oleh orang-orang, Facebook bisa melihat apa yang mereka bagikan, dan Amazon mengetahui apa yang mereka beli. Mereka seolah memiliki “God’s eyes view” tentang aktivitas di pasar mereka sendiri dan sekitarnya.

Luar biasa ya? Sekarang hampir semua perusahaan mengumpulkan data untuk sumber daya mereka.

Lantas, bagaimana cara mengumpulkan data? Ada tiga cara yang bisa kita lakukan untuk mengumpulkan data, yaitu.

* Mengekstrasi data (misal dari internet, riset, survei, dll).
* Mengumpulkan dan membuat dataset Anda sendiri dari nol.
* Menggunakan dataset yang telah ada.

Untuk saat ini, kita akan menggunakan dataset yang sudah ada dari platform penyedia data. Di masa mendatang tentu Anda dapat mencoba mengekstrak atau mengumpulkan dataset Anda sendiri ya.

Menemukan dataset yang tepat adalah salah satu langkah penting dalam proyek machine learning. Saat ini, tersedia banyak sumber data di internet yang dapat kita manfaatkan. Beberapa di antaranya yang perlu Anda ketahui adalah sebagai berikut.

* UC Irvine Machine Learning Repository

[UCI ML Repository](https://archive.ics.uci.edu/datasets) adalah kumpulan database, teori, dan generator data yang digunakan oleh komunitas ML untuk analisis algoritma machine learning. Arsip tersebut awalnya dibuat sebagai arsip ftp pada tahun 1987 oleh David Aha, seorang mahasiswa pascasarjana UC Irvine. Sejak saat itu database UCI ML Repository ini digunakan secara luas oleh mahasiswa, staf pengajar, dan peneliti di seluruh dunia sebagai salah satu sumber utama dataset machine learning.

* Kaggle Dataset

[Kaggle](https://www.kaggle.com/datasets) adalah komunitas belajar ilmu data paling populer di dunia. Kaggle memiliki peralatan dan sumber daya yang kuat untuk membantu kita belajar data science dan machine learning. Saat ini Kaggle memiliki 50.000 lebih publik dataset, baik dataset bersifat dummy ataupun riil yang dapat Anda unduh secara bebas.

* Google Dataset Search Engine

Pada akhir tahun 2018 Google meluncurkan [Dataset Search](https://datasetsearch.research.google.com/), sebuah mesin pencari dataset. Tools ini bertujuan untuk menyatukan ribuan repositori dataset yang berbeda agar dataset tersebut lebih mudah ditemukan oleh pengguna.

* Tensorflow Dataset

Seperti yang telah dijelaskan pada sub-modul sebelumnya, [TensorFlow](https://www.tensorflow.org/) adalah framework open source untuk machine learning yang dikembangkan dan digunakan oleh Google. Selain menyediakan [learning resources](https://www.tensorflow.org/learn), tensorflow juga menyediakan [data resources](https://www.tensorflow.org/datasets/catalog/overview) yang cukup lengkap di library-nya mulai dari audio data, images, text, video, dan lainnya.

* US Government Data

Bagi Anda yang tertarik untuk mempelajari fenomena yang terjadi di Amerika Serikat, pemerintah Amerika meluncurkan [data online resources](https://www.data.gov/) yang mudah diakses oleh publik. Isinya antara lain data badai, data angka kelulusan dan dropout, data hewan-hewan yang terancam punah, statistik kriminal, dan berbagai data menarik lainnya.

* Satu Data Indonesia

Pemerintah Indonesia, melalui portal resmi [Satu Data Indonesia](https://data.go.id/) menjalankan kebijakan tata kelola data pemerintah yang bertujuan untuk menciptakan data berkualitas, mudah diakses, dapat dibagi, dan digunakan oleh Instansi Pusat serta Daerah. Data dalam portal ini dapat diakses secara terbuka dan dikategorikan sebagai data publik, sehingga tidak memuat rahasia negara, rahasia pribadi, atau hal lain sejenisnya sebagaimana diatur dalam Undang-undang nomor 14 Tahun 2008 tentang Keterbukaan Informasi Publik.

* Open Data Pemerintah Jawa Barat

[Open data Jawa Barat](https://opendata.jabarprov.go.id/id/dataset)adalah portal resmi data terbuka milik Pemerintah Provinsi Jawa Barat yang berisikan data-data dari Perangkat Daerah di lingkungan Pemerintah Provinsi Jawa Barat. Open Data Jawa Barat ada untuk dapat memenuhi kebutuhan data publik bagi masyarakat. Data disajikan dengan akurat, akuntabel, valid, mudah diakses dan berkelanjutan.