MACHINE LEARNING

Feature Engineering & Integrasi Data

Magister Teknik Informatika

Dr. Chairani, S.Kom., M.Eng
IIB DARMAJAYA, 2023/2024

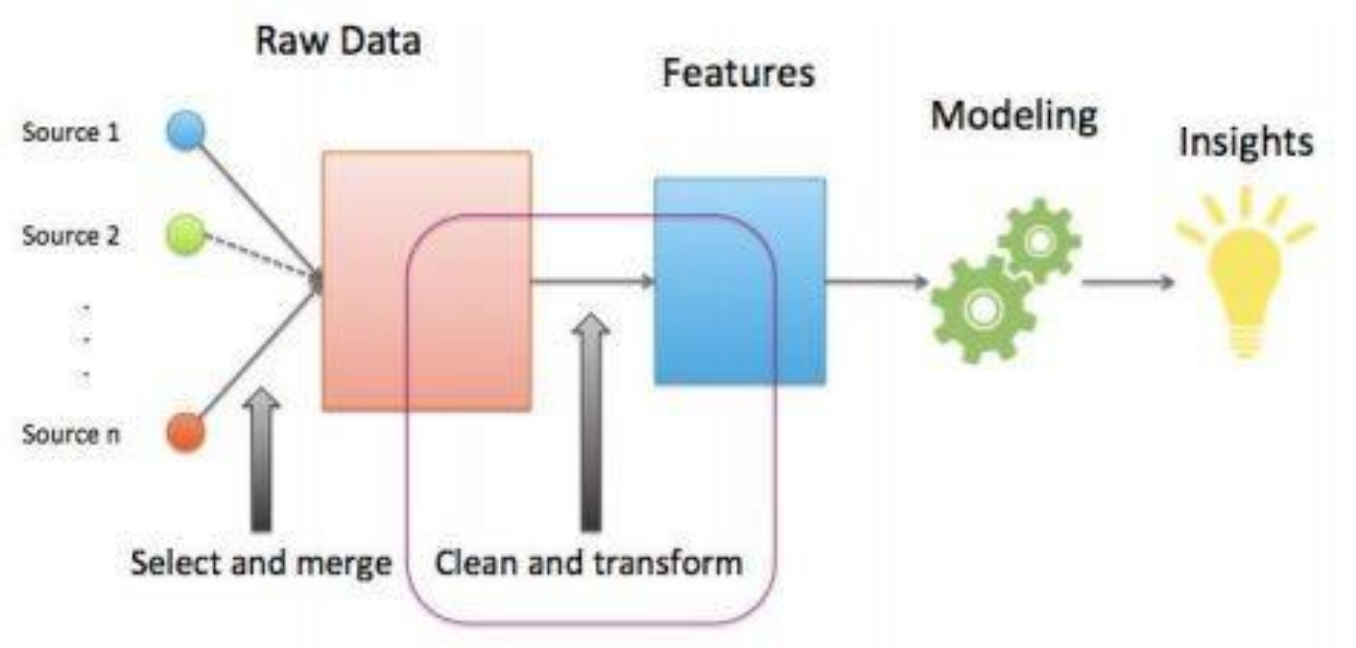
Subject

- Bagian dari Data Preparation
- Data Preparation yang dibahas adalah transformasi data yaitu:
  - Rekayasa Fitur
  - Pelabelan data
- Ada beberapa teknik transformasi data yang digunakan sesuai kebutuhan dan ketersediaan/jenis data baik numerik maupun kategorik

Outline

- Dokumentasi Rekayasa Fitur
- Pelabelan Data
- Dokumentasi Proses Pelabelan Data
- Integrasi Data
- Sesi Hands On

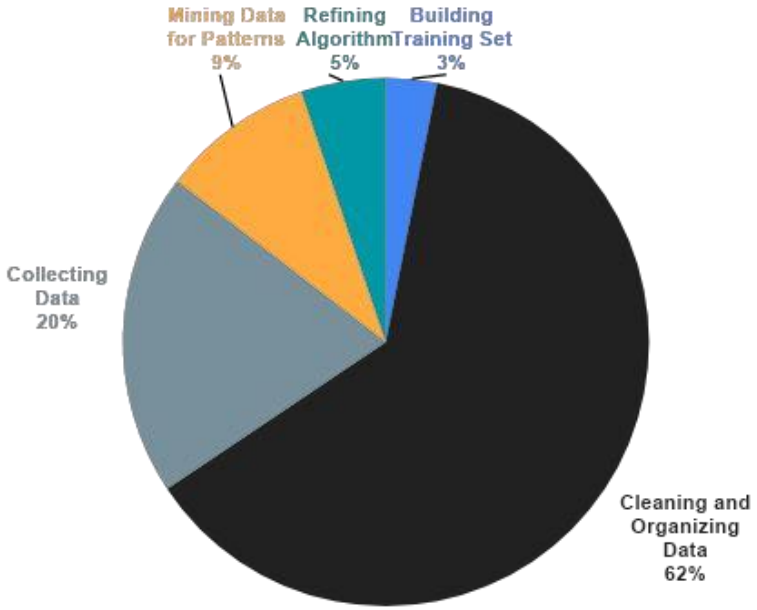
Proses Pembentukan Model



Why Feature Engineering Matters?

Data scientists and machine learning engineers frequently **gather data** in order **to solve a real-life problem.** These engineers have the unique job of engineering pipelines and architectures **designed to handle and transform raw data into something usable** by the rest of the company, particularly the data scientists and machine learning engineers.

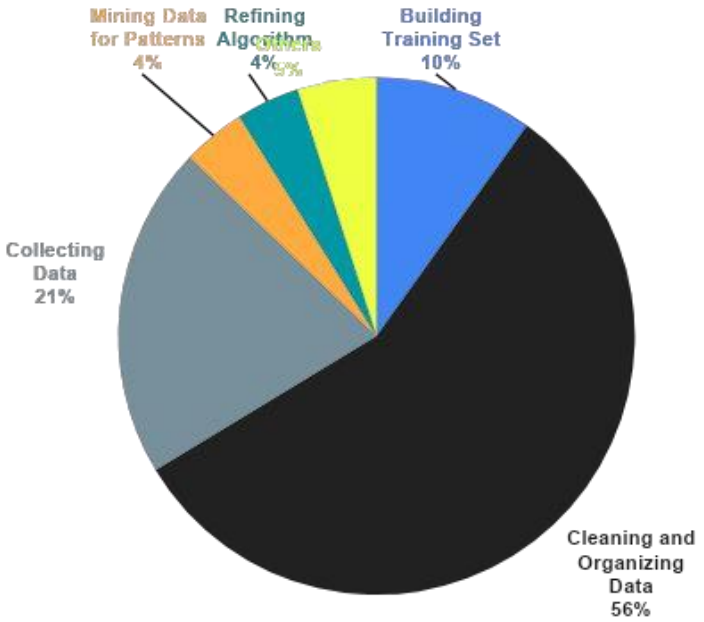
Why Feature Engineering Matters?



Sebuah survei yang dilakukan oleh para ilmuwan data di lapangan mengungkapkan bahwa lebih dari **80% waktu** dari data scientist dihabiskan untuk mengumpulkan, membersihkan, dan mengorganisir data. **Kurang dari 20%** sisa waktu mereka dihabiskan untuk membangun algoritma atau model.

[1]https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/.

Why Feature Engineering Matters?



Sebuah survei yang sama juga dilakukan untuk mengetahui bagian pekerjaan mana yang dirasa kurang menyenangkan, hasilnya **77% responden** mengatakan fase mengumpulkan, membersihkan, dan mengorganisir data. adalah fase yang dirasa **kurang menyenangkan**.

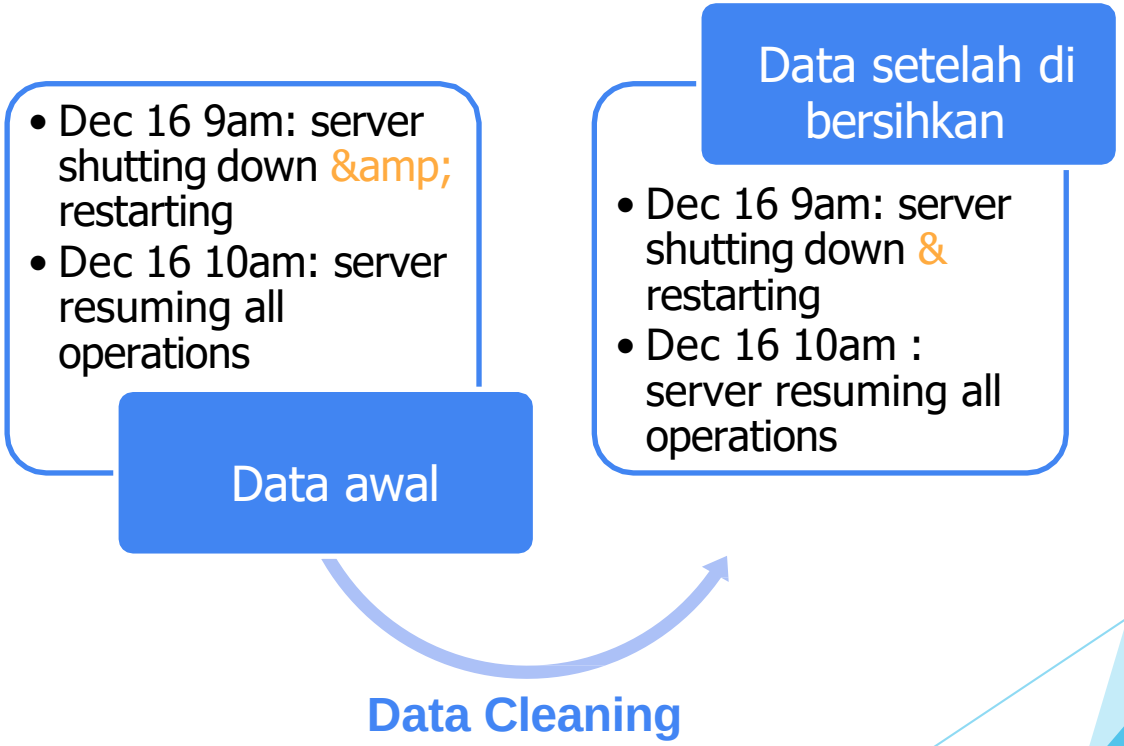
[1]https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/.

Why Feature Engineering Matters

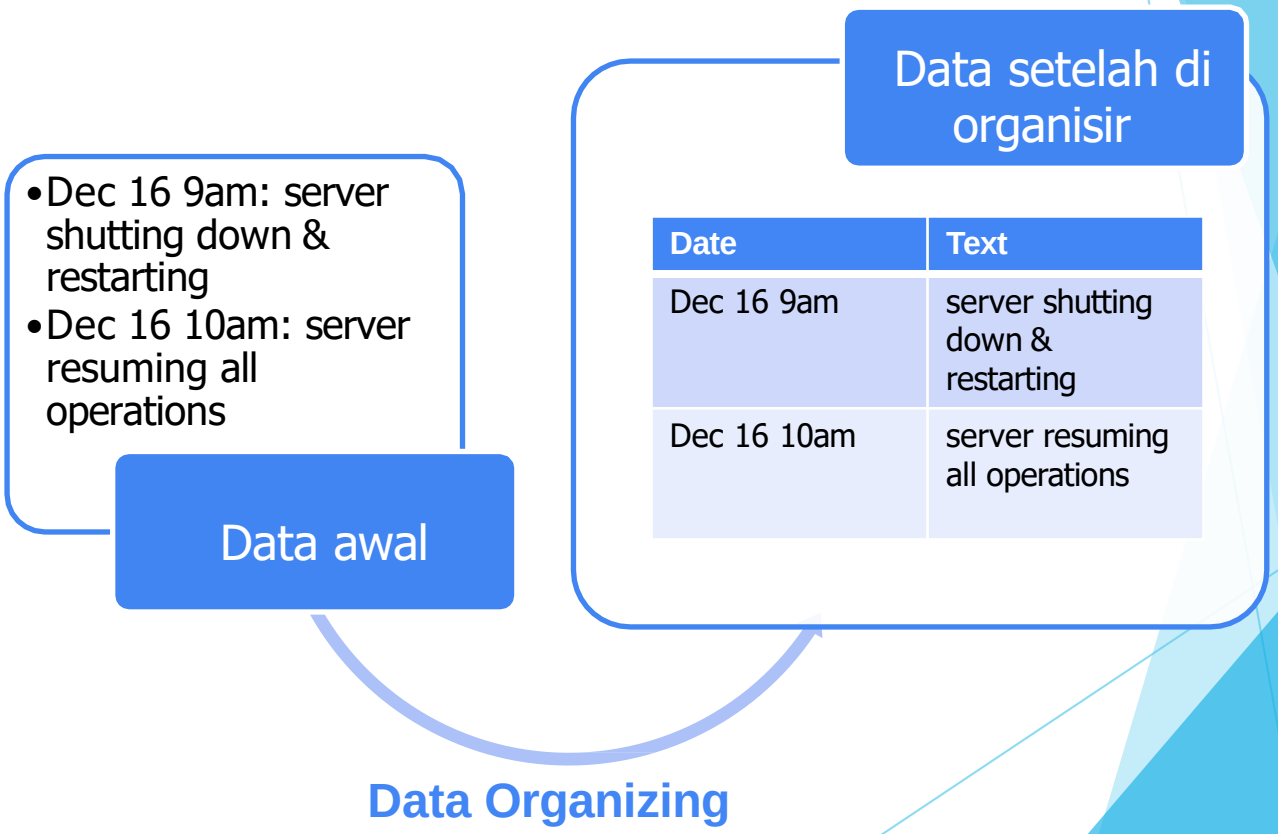
"A stellar data scientist knows that *preparing data is not only so important that it takes up most of their time*, they also know that it is an arduous process and can be unenjoyable. **Far too often, we take for granted clean data given to us by machine learning competitions and academic sources**. More than 90% of data, the data that is interesting, and the most useful, exists in this raw format."

Dikutip dari: Sinan Ozdemir. "Feature Engineering Made Easy."

Data Cleaning



Data Organizing



Dokumentasi Fitur

Sebagian besar **Data Scientist**

cenderung melakukan

proses dokumentasi

di **'kepala'** mereka.

Perlunya Dokumentasi Data/Fitur

Dokumentasi data dapat **menjembatani kesenjangan** antara **transaksi** (pembuatan data) dan **analisis** (konsumsi data). Dokumentasi data yang baik memungkinkan pengguna, ataupun rekan tim untuk memahami siapa/apa/kapan/di mana/bagaimana/mengapa data tersebut dibentuk ataupun dikonsumsi.

Paramater/Daftar Isi Dokumentasi Data Transfomation

Laporan dokumentasi data cleaning, setidaknya memiliki parameter berikut:

- Fitur awal dan rekayasa fitur yang digunakan
- Teknik transformasi data yang diterapkan
  - Apakah algoritma pemodelan mengharapkan jenis data tertentu, seperti numerik? Jika demikian, lakukan transformasi yang diperlukan
  - Apakah data perlu dinormalisasi sebelum pemodelan?
  - Bisakah atribut yang hilang dibangun menggunakan agregasi, rata-rata, atau induksi?
- Hasil transformasi
- Rekomendasi transformasi

Pelabelan Data

Pelabelan Data - Intro

- **Kuantitas & kualitas data pelatihan** yang secara langsung **menentukan keberhasilan** suatu **algoritma AI** sehingga tidak mengherankan jika rata-rata 80% waktu yang dihabiskan untuk proyek AI membahas data pelatihan yang mencakup proses **pelabelan data**.
- **Keakuratan model AI Anda berkorelasi langsung dengan kualitas data** yang digunakan untuk melatihnya.
- Hal ini menjadi satu alasan mengapa ***proses pelabelan data merupakan bagian integral dari alur kerja persiapan data dalam membangun model AI yang andal***.

Pelabelan Data - Intro

Pelabelan data dalam konteks pembelajaran mesin adalah proses mendeteksi serta menandai sampel data.

- Tahapan proses ini menjadi sangat penting dalam hal pembangunan model dengan pendekatan pembelajaran mesin berbasiskan supervised-learning.
- Pelabelan Data mengacu pada proses menambahkan tag atau label pada data masukan yang berbentuk *gambar*, *video*, *teks*, dan *audio*.
- **Tag** ini **membentuk representasi** dari kelas **objek** apa yang dimiliki data dan membantu model pembelajaran mesin untuk mengidentifikasi kelas objek tertentu saat ditemui dalam data tanpa tag.
- Secara umum, pelabelan data dapat merujuk pada tugas yang mencakup **penandaan data**, **anotasi**, **klasifikasi**, **moderasi**, **transkripsi**, atau **pemrosesan**.

Pelabelan Data - Intro

Pelabelan data adalah bagian utama dari alur kerja pra pemrosesan data untuk machine learning. Data berlabel ini kemudian digunakan untuk melatih model pembelajaran mesin untuk menemukan "makna" dalam data baru yang serupa dan relevan.

Pelabelan data menyusun data untuk membuatnya bermakna.

Sepanjang proses ini, data scientist berusaha keras untuk memperoleh kualitas dan kuantitas yang baik.

Label yang lebih akurat ditambah dengan jumlah data berlabel yang lebih besar menciptakan model pembelajaran mendalam yang lebih berguna,

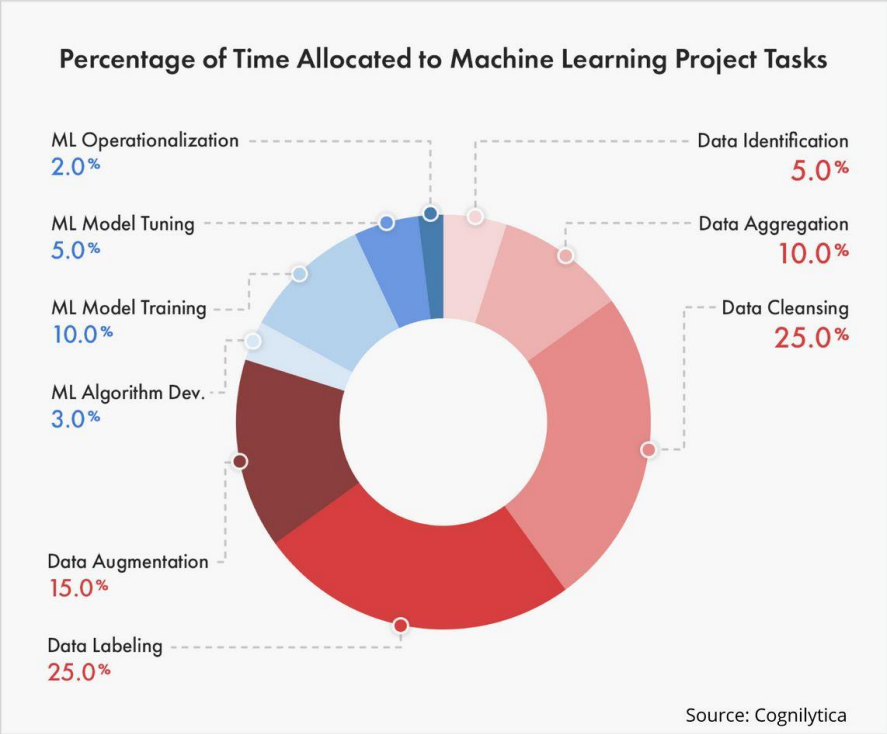
karena model pembelajaran mesin yang dihasilkan mendasarkan keputusan mereka pada semua data berlabel.

Pelabelan Data - Data Training

Data pelatihan mengacu pada data yang telah dikumpulkan untuk diumpankan ke model pembelajaran mesin untuk membantu model mempelajari data lebih lanjut.

- Data pelatihan dapat berupa berbagai bentuk, termasuk gambar, suara, teks, atau fitur tergantung pada model pembelajaran mesin yang digunakan dan tugas ataupun business goal yang ingin dicapai.
- Data pelatihan bisa diberi anotasi maupun tidak diberi anotasi.
- Ketika data pelatihan dianotasi, label tersebut disebut sebagai dasar kebenaran atau ***Ground Truth - istilah digunakan untuk informasi yang telah diketahui sebelumnya bernilai benar***.

Pelabelan Data - Mengapa Proses Pelabelan Diperlukan ?



- Anda memiliki banyak data yang tidak berlabel.
- Sebagian besar data tidak dalam bentuk berlabel, hal merupakan tantangan bagi sebagian besar tim proyek Data Science.
- Sepenuhnya 80% dari waktu proyek berbasis AI dihabiskan untuk mengumpulkan, mengatur, dan memberi label data,
- menurut firma analis Cognilytica, dan ini adalah waktu yang tidak dapat dihabiskan oleh tim karena mereka berlomba untuk mendapatkan data yang dapat digunakan, yaitu data yang terstruktur dan diberi label dengan benar untuk melatih dan menerapkan model.

Pelabelan Data - Unlabelled vs Labelled

- Dataset pelatihan bergantung pada jenis permasalahan ataupun tujuan model pembelajaran mesin yang ingin kita bentuk.
- Algoritma Machine/Deep Learning dapat secara luas diklasifikasikan berdasarkan jenis data yang mereka butuhkan dalam tiga tipe, yaitu :
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning

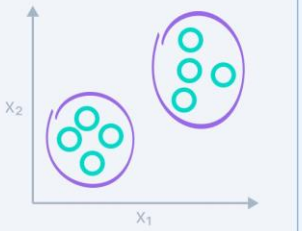
Pelabelan Data - Supervised Learning

- Pembelajaran terawasi, jenis yang paling umum, adalah jenis algoritmea pembelajaran mesin yang memerlukan data dan label yang sesuai untuk dilatih.
- Pendekatan ini biasanya digunakan untuk menyelesaikan permasalahan **klasifikasi** dan **segmentasi**.
- Prosedur pelatihan tipikal terdiri dari memasukkan data yang telah beranotasi ke mesin untuk membantu model belajar, dan kemudian melakukan pengujian model yang terbentuk pada data yang tak beranotasi.
- Untuk menemukan keakuratan metode tersebut, data beranotasi (ground truth) dengan label tersembunyi biasanya digunakan dalam tahap pengujian algoritma.
- Dengan demikian, data beranotasi merupakan kebutuhan mutlak untuk melatih model pembelajaran mesin secara terawasi.

Pelabelan Data - Unsupervised Learning

- Dalam pembelajaran tanpa pengawasan, data input merupakan data tanpa anotasi dan model berlatih tanpa pengetahuan tentang label yang mungkin dimiliki data input.
- Algoritma unsupervised termasuk autoencoder yang memiliki output yang sama dengan inputnya.
- Metode pembelajaran tanpa pengawasan juga mencakup algoritma pengelompokan yang mengelompokkan data ke dalam cluster 'n', di mana 'n' adalah hyperparameter.

Pelabelan Data - Supervised vs Unsupervised

Supervised learning	Unsupervised learning
Input data is labeled	Input data is unlabeled
Has a feedback mechanism	Has no feedback mechanism
Data is classified based on the training dataset	Assigns properties of given data to classify it
Divided into Regression & Classification	Divided into Clustering & Association
Used for prediction	Used for analysis
Algorithms include: decision trees, logistic regressions, support vector machine	Algorithms include: k-means clustering, hierarchical clustering, apriori algorithm
A known number of classes	A unknown number of classes
	

V7 Labs

Pelabelan Data - Semi-supervised Learning

- Dalam pembelajaran semi-diawasi, kombinasi data beranotasi dan tidak beranotasi digunakan untuk melatih model.
- Meskipun hal ini mengurangi biaya anotasi data dengan menggunakan kedua jenis data tersebut, pada umumnya pendekatan ini menggunakan banyak asumsi pada data pelatihan yang digunakan untuk membangun model.
- Kasus penggunaan pembelajaran semi-diawasi salah satunya klasifikasi urutan Protein dan analisis konten Internet.

Pelabelan Data - Semi-supervised Learning

Dataset awal :

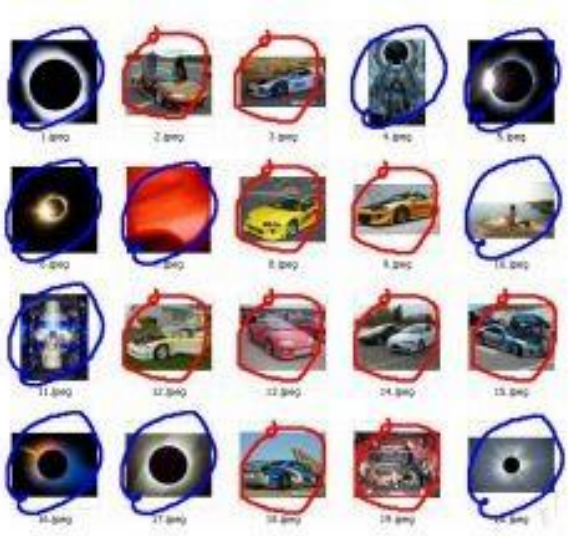
Eclipse

Non- eclipse

Dataset yang telah diperkaya :

Pelabelan Data - Semi-supervised Learning

Dataset yang telah dilabeli secara manual



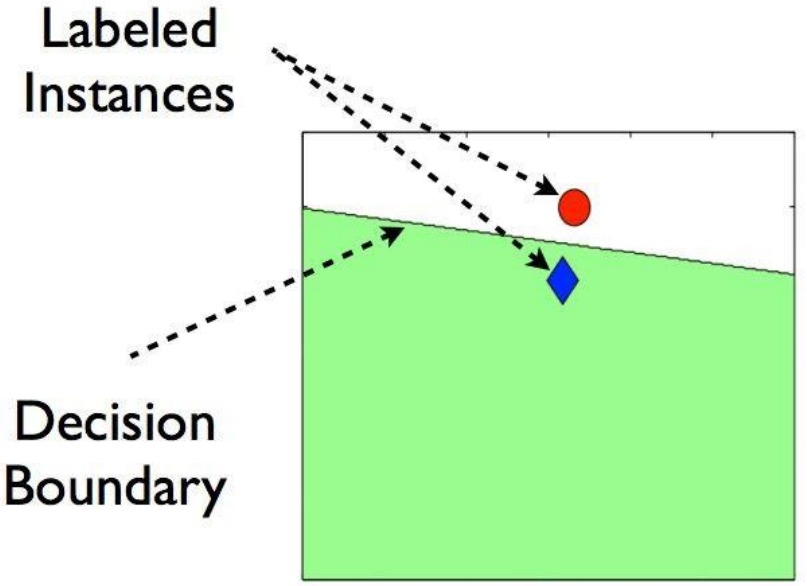
Setelah menjalankan algoritma supervised learning pada studi kasus ini, hasil model kedua **pasti** mengungguli model pertama yang dibangun hanya berisi dua gambar sebagai data pelatihan.

Tetapi pendekatan ini hanya berlaku untuk tujuan kecil karena anotasi/pelabelan manual ke kumpulan data besar bisa sangatlah sulit dan mahal.

untuk memecahkan jenis masalah ini, terdapat jenis pembelajaran yang berbeda yang dikenal sebagai ***semi-supervised learning***, yang menggunakan baik data berlabel (pembelajaran terbimbing) dan data tidak berlabel (pembelajaran tanpa pengawasan).

Pelabelan Data - Peran data tanpa label

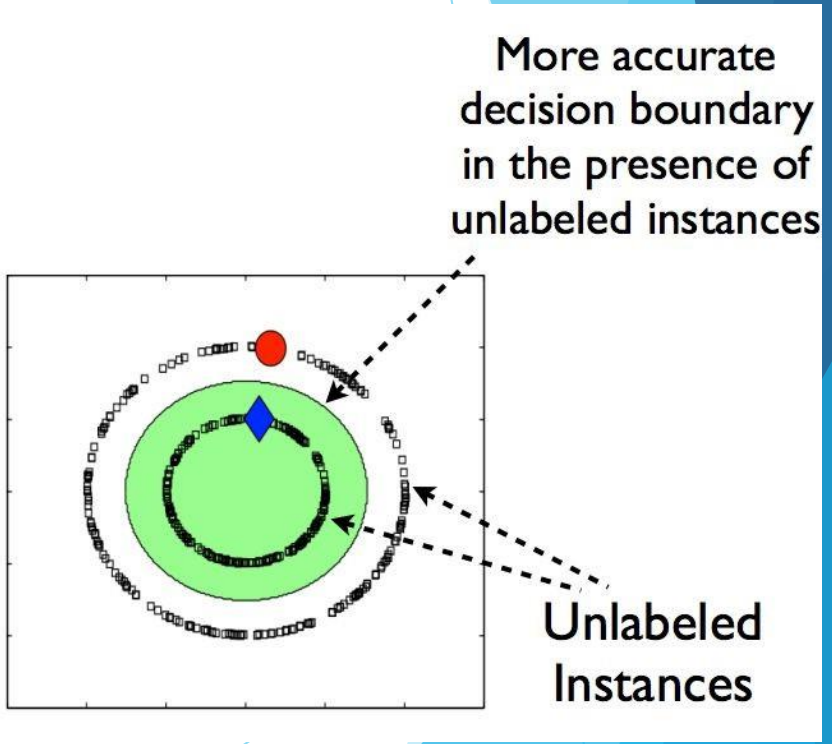
Pada pembangunan model dengan data berlabel, Anda hanya memiliki dua titik data yang termasuk dalam dua kategori berbeda, dan garis yang ditarik adalah batas keputusan dari setiap model yang diawasi.



Pelabelan Data - Peran data tanpa label

Selanjutnya, katakanlah kita menambahkan beberapa data yang tidak berlabel ke data ini seperti yang ditunjukkan pada gambar samping.

Gambar ini (di kanan), garis pembatas antara wilayah hijau dan putih menjadi lebih presisi. Perbedaan garis pembatas tersebut menunjukkan bahwa dengan memberikan menambahkan data yang tidak berlabel, garis batas keputusan model menjadi lebih akurat.



Pelabelan Data - Peran data tanpa label

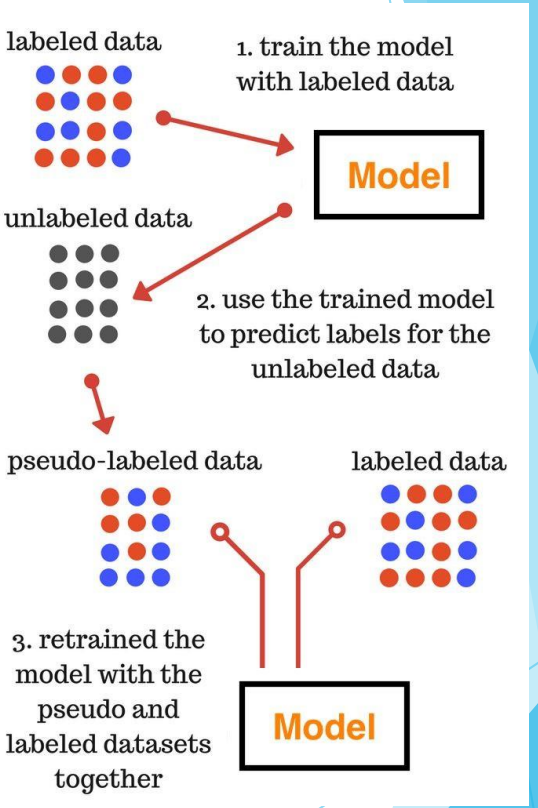
Contoh kasus

Keuntungan menggunakan data tidak berlabel adalah:

- Data berlabel mahal dan sulit didapat sedangkan data tidak berlabel berlimpah dan murah.
- Meningkatkan ketahanan model dengan batas keputusan yang lebih tepat.

Pseudo Labelling

- Manusia tidak hanya belajar dari informasi tetapi mampu memahami suatu berdasarkan kesamaan karakteristik yang dimiliki
- Bisakah kita membangun sistem yang membutuhkan pengawasan minimal yang dapat mempelajari sebagian besar tugas sendiri?
- Terdapat berbagai teknik penerapan semi-Supervised Learning, salah satu tekniknya adalah Teknik Pelabelan Pseudo atau pseudo-labelling.



Ref :
4 Q. Xie, M.-T. Luong, E. Hovy, and Q.V. Le, "Self-training with noisy student improves ImageNet classification," arXiv:1911.04252, 2020.
5 I.Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," arXiv:1905.00546, 2019.
6 K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E.D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," arXiv:2001.07685, 2020.

Human in The Loop (HITL)

Istilah Human-In-The-Loop paling sering mengacu pada pengawasan konstan dan validasi hasil model AI oleh manusia.

Ada dua cara utama di mana manusia menjadi bagian dari loop Machine Learning:

- Memberi label pada data pelatihan: Seorang anotator (*experts*) diwajibkan untuk memberi label pada data pelatihan yang diumpankan ke model pembelajaran mesin (diawasi/semi-diawasi).
- Melatih model: Data scientist melatih model dengan terus-menerus mengawasi detail model seperti fungsi kerugian (*loss function*) dan hasil prediksi.
- Terkadang kinerja model dan prediksi divalidasi oleh manusia dan hasil validasi diumpankan kembali ke model.

Pendekatan Pelabelan Data

- Pendekatan pelabelan bergantung pada pernyataan masalah, kerangka waktu proyek, dan jumlah orang yang terkait dengan pekerjaan.
- Pelabelan internal dan crowdsourcing sangat umum, terminologi ini juga dapat mencakup bentuk-bentuk baru pelabelan baru dan anotasi yang memanfaatkan AI dan pembelajaran aktif (*active learning*) untuk melakukan tugas pelabelan/anotasi tersebut.
- Pendekatan yang paling umum untuk anotasi data tercantum di bawah ini
    - In-house data labelling
    - Crowdsourcing
    - Outsourcing
    - Machine-based annotation

In-house Labelling

- Memiliki kualitas tertinggi dan umumnya dilakukan oleh data scientist dan insinyur data yang dipekerjakan di organisasi.
- Pelabelan berkualitas tinggi sangat penting untuk industri seperti asuransi atau perawatan kesehatan, dan seringkali memerlukan konsultasi dengan para ahli di bidang terkait untuk pelabelan data yang tepat.
- Seperti yang diharapkan untuk pelabelan internal, dengan peningkatan kualitas anotasi, waktu yang dibutuhkan untuk membuat anotasi cukup tinggi, sehingga seluruh proses pelabelan dan pembersihan data menjadi sangat lambat.

Crowdsourcing

- Crowdsourcing mengacu pada proses memperoleh data beranotasi dengan bantuan sejumlah besar pekerja lepas yang terdaftar di platform crowdsourcing .
- Kumpulan data yang dianotasi sebagian besar terdiri dari data sepele seperti gambar hewan, tumbuhan, dan lingkungan alam dan tidak memerlukan keahlian tambahan.
- Oleh karena itu, tugas membuat anotasi pada kumpulan data sederhana sering kali dilakukan secara crowdsource ke platform yang memiliki puluhan ribu annotator data terdaftar.

Outsourcing Labelling

- Outsourcing adalah jalan tengah antara crowdsourcing dan pelabelan data internal di mana tugas anotasi data dialihdayakan ke organisasi atau individu.
- Salah satu keuntungan outsourcing adalah kita dapat menilai topik tertentu sebelum pekerjaan diserahkan.
- Pendekatan membangun kumpulan data anotasi ini sangat cocok untuk proyek yang tidak memiliki banyak dana, namun membutuhkan kualitas anotasi data yang signifikan.

Machine-based Annotation

- Salah satu bentuk anotasi yang paling baru adalah anotasi berbasis mesin.
- Anotasi berbasis mesin mengacu pada penggunaan alat anotasi dan otomatisasi yang secara drastis dapat meningkatkan kecepatan anotasi data tanpa mengorbankan kualitas hasil pelabelan.
- Perkembangan otomatisasi baru-baru ini dalam alat anotasi mesin tradisional—menggunakan algoritma pembelajaran mesin yang tidak diawasi dan semi-diawasi—membantu secara signifikan mengurangi beban kerja pada pemberi label manusia.

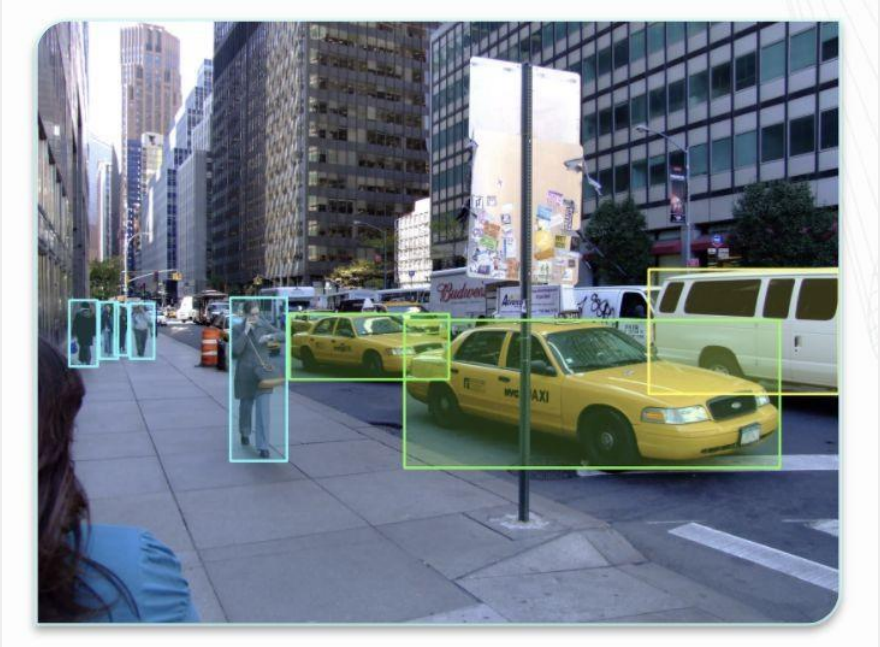
- Algoritma tanpa pengawasan (***unsupervised learning***) seperti pengelompokan dan algoritma ***semi-***

Kelebihan dan Kekurangan Masing-masing Metode

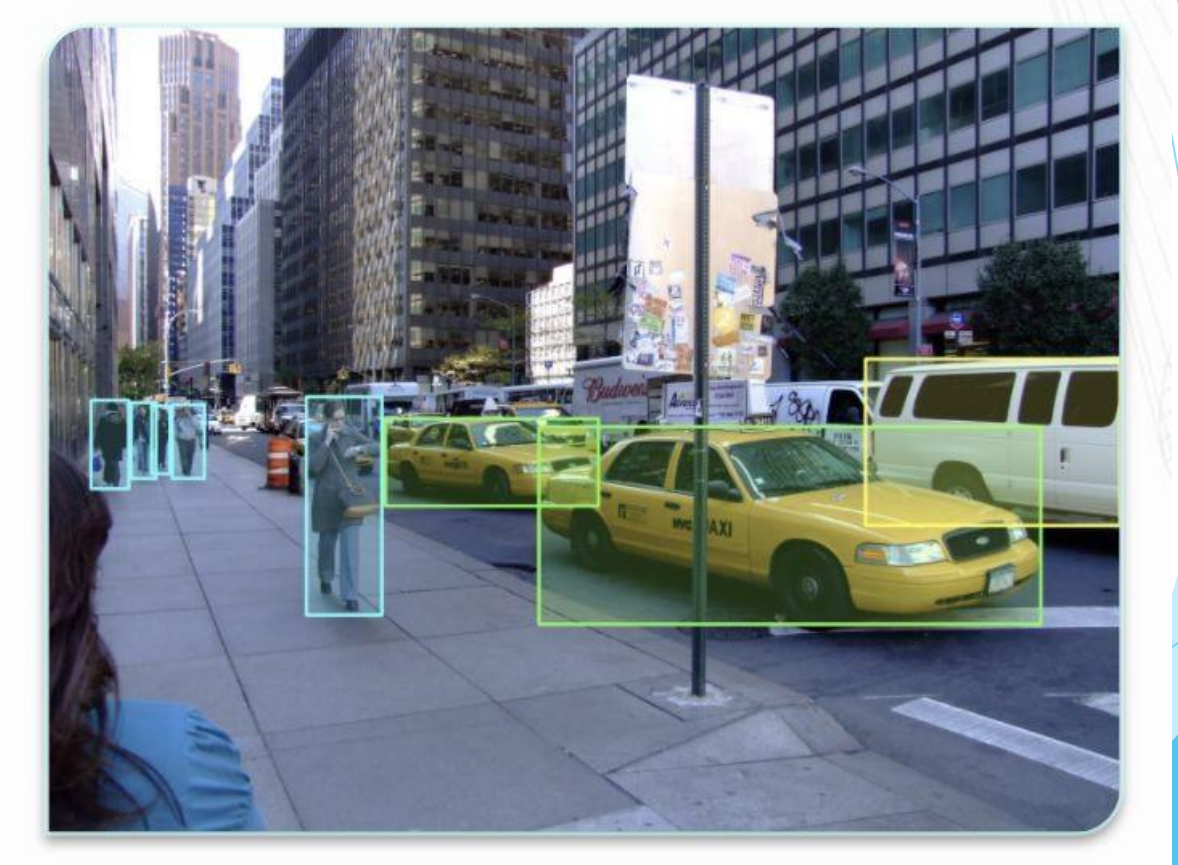
Approach	Description	Pros	Cons
Internal labeling	Assignment of tasks to an in-house data science team	✓ Predictable results ✓ High accuracy of labeled data ✓ The ability to track progress	× It takes much time
Outsourcing	Recruitment of temporary employees on freelance platforms, posting vacancies on social media and job search sites	✓ The ability to evaluate applicants' skills	× The need to organize workflow
Crowdsourcing	Cooperation with freelancers from crowdsourcing platforms	✓ Cost savings ✓ Fast results	× Quality of work can suffer
Specialized outsourcing companies	Hiring an external team for a specific project	✓ Assured quality	× Higher price compared to crowdsourcing
Synthetic labeling	Generating data with the same attributes of real data	✓ Fewer constraints for using sensitive and regulated data ✓ Training data without mismatches and gaps ✓ Cost- and time-effectiveness	× High computational power required
Data programming	Using scripts that programmatically label data to avoid manual work	✓ Automation ✓ Fast results	× Lower quality dataset

altexsoft
software r&d engineering

Pelabelan data : Pengolahan Citra



Pelabelan data : Pengolahan Citra



Pelabelan data : Pengolahan Citra

Permasalahan pengolahan citra memerlukan data visual beranotasi dalam bentuk gambar. Anotasi data dalam pengolahan citra bergantung pada tugas visual yang kita inginkan untuk dilakukan oleh model.

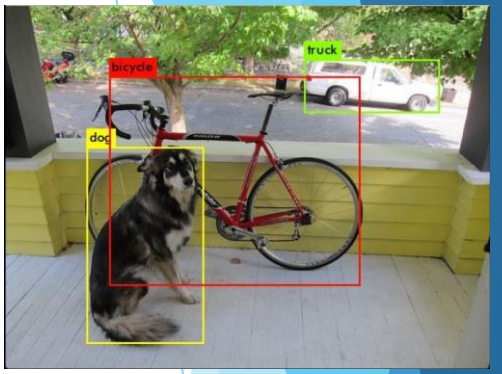
Jenis anotasi data umum pada kasus pengolahan citra antara lain :

- *Image Classification* - klasifikasi gambar
  - Anotasi data untuk klasifikasi gambar memerlukan penambahan tag ke gambar yang sedang dikerjakan.
  - Jumlah tag unik di seluruh database adalah jumlah kelas yang dapat diklasifikasi oleh model.
  - Masalah klasifikasi dapat dibagi lagi menjadi:
    - Klasifikasi kelas biner (yang hanya terdiri dari dua tag)
    - Klasifikasi multiclass (yang berisi beberapa tag)
  - Selain itu, klasifikasi multi-label juga dapat dilihat, terutama dalam hal deteksi penyakit, dan mengacu pada setiap gambar yang memiliki lebih dari satu tag.
- *Image Segmentation* - segmentasi gambar
  - Dalam Segmentasi Gambar, tugas algoritma pengolahan citra (*Computer Vision*) adalah memisahkan objek dalam gambar dari latar belakangnya dan objek lain dalam gambar yang sama.
  - Ini umumnya berarti peta piksel dengan ukuran yang sama dengan gambar yang berisi 1 di mana objek ada dan 0 di mana anotasi belum dibuat.
  - Untuk beberapa objek yang akan disegmentasi dalam gambar yang sama, peta piksel untuk setiap objek digabungkan berdasarkan saluran dan digunakan sebagai kebenaran dasar untuk model.

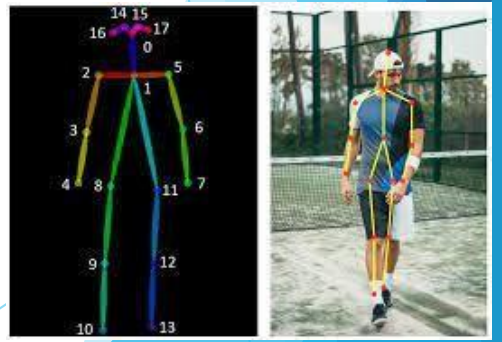
Sesi Lab 10.3

Pelabelan data : Pengolahan Citra

- *Object Detection* - Deteksi obyek
  - Deteksi Objek mengacu pada deteksi objek dan lokasinya melalui pengolahan citra.
  - Anotasi data dalam deteksi objek sangat berbeda dari yang ada di Klasifikasi Gambar, dengan setiap objek dianotasi menggunakan kotak pembatas (*bounding box*).
  - Kotak pembatas adalah segmen persegi panjang terkecil yang berisi objek dalam gambar.
  - Anotasi kotak pembatas biasanya disertai dengan tag di mana setiap kotak pembatas diberi label pada gambar.
  - Umumnya, koordinat kotak pembatas ini dan tag yang sesuai untuknya disimpan dalam file JSON terpisah dalam format kamus dengan nomor gambar/ID gambar menjadi kunci kamus.



- *Pose Estimation* - Estimasi pose
  - Estimasi pose mengacu pada penggunaan alat Computer Vision untuk memperkirakan pose seseorang dalam sebuah gambar.
  - Estimasi pose berjalan dengan mendeteksi titik-titik kunci dalam tubuh dan menghubungkan titik-titik kunci ini untuk mendapatkan pose.
  - *Ground Truth* (GT) yang sesuai untuk model estimasi pose, menjadi poin kunci dari sebuah gambar.
  - *Ground Truth* (GT) dapat berupa data koordinat sederhana yang diberi label dengan bantuan tag, di mana setiap koordinat memberikan lokasi titik kunci tertentu, yang diidentifikasi oleh tag, pada gambar masing-masing.



Pelabelan data : Natural Language Processing (NLP)

Pelabelan data : Natural Language Processing (NLP)

Pemrosesan bahasa alami (atau disingkat NLP) mengacu pada analisis bahasa manusia dan bentuknya selama interaksi baik dengan manusia lain maupun dengan mesin. Menjadi bagian dari linguistik komputasi awalnya, NLP telah berkembang lebih lanjut dengan bantuan Artificial Intelligence dan Deep Learning.

Berikut adalah beberapa pendekatan pelabelan data untuk pelabelan data NLP :

- Entity annotation and linking
  - Anotasi entitas mengacu pada anotasi entitas atau fitur tertentu dalam korpus data yang tidak berlabel.
  - Kata 'Entitas' dapat mengambil bentuk yang berbeda tergantung pada tugas yang dihadapi.
  - Untuk anotasi kata benda yang tepat, kami telah menamai anotasi entitas yang mengacu pada identifikasi dan penandaan nama dalam teks.
  - Untuk analisis frasa, kami mengacu pada proses sebagai penandaan frasa kunci di mana kata kunci atau frasa kunci dari teks dianotasi.
  - Untuk analisis dan anotasi elemen fungsional dari teks apapun seperti kata kerja, kata benda, preposisi, kami menggunakan penandaan Parts of Speech, disingkat sebagai penandaan POS.
  - Penandaan POS digunakan dalam penguraian, terjemahan mesin, dan pembuatan data linguistik.
  - Anotasi entitas diikuti dengan penautan entitas, di mana entitas beranotasi ditautkan ke repositori data di sekitarnya untuk menetapkan identitas unik ke masing-masing entitas ini. Hal ini sangat penting ketika teks berisi data yang dapat ambigu dan perlu disambiguasi.
  - Tautan entitas sering digunakan untuk anotasi semantik, di mana informasi semantik entitas ditambahkan sebagai anotasi.

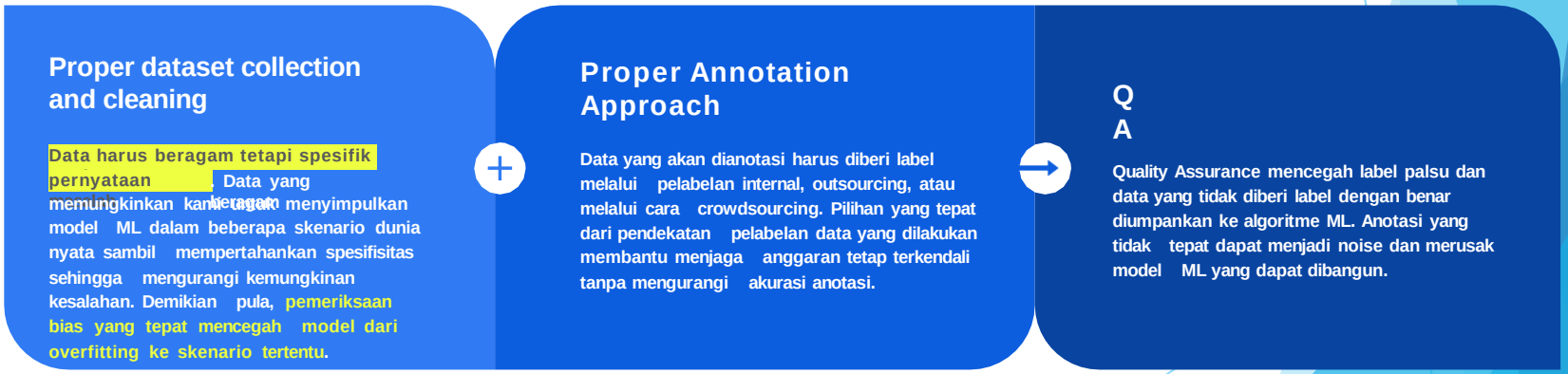
Pelabelan data : Natural Language Processing (NLP)

- Text Classification
  - Mirip dengan klasifikasi gambar di mana kami menetapkan label ke data gambar, dalam klasifikasi teks, menetapkan satu atau beberapa label ke blok teks.
  - Sementara dalam anotasi dan penautan entitas, kita memisahkan entitas di dalam setiap baris teks, dalam klasifikasi teks, teks dianggap sebagai keseluruhan dan satu set tag ditetapkan ke dalamnya
  - Jenis klasifikasi teks meliputi klasifikasi berdasarkan sentimen (untuk analisis sentimen) dan klasifikasi berdasarkan topik yang ingin disampaikan teks (untuk kategorisasi topik).

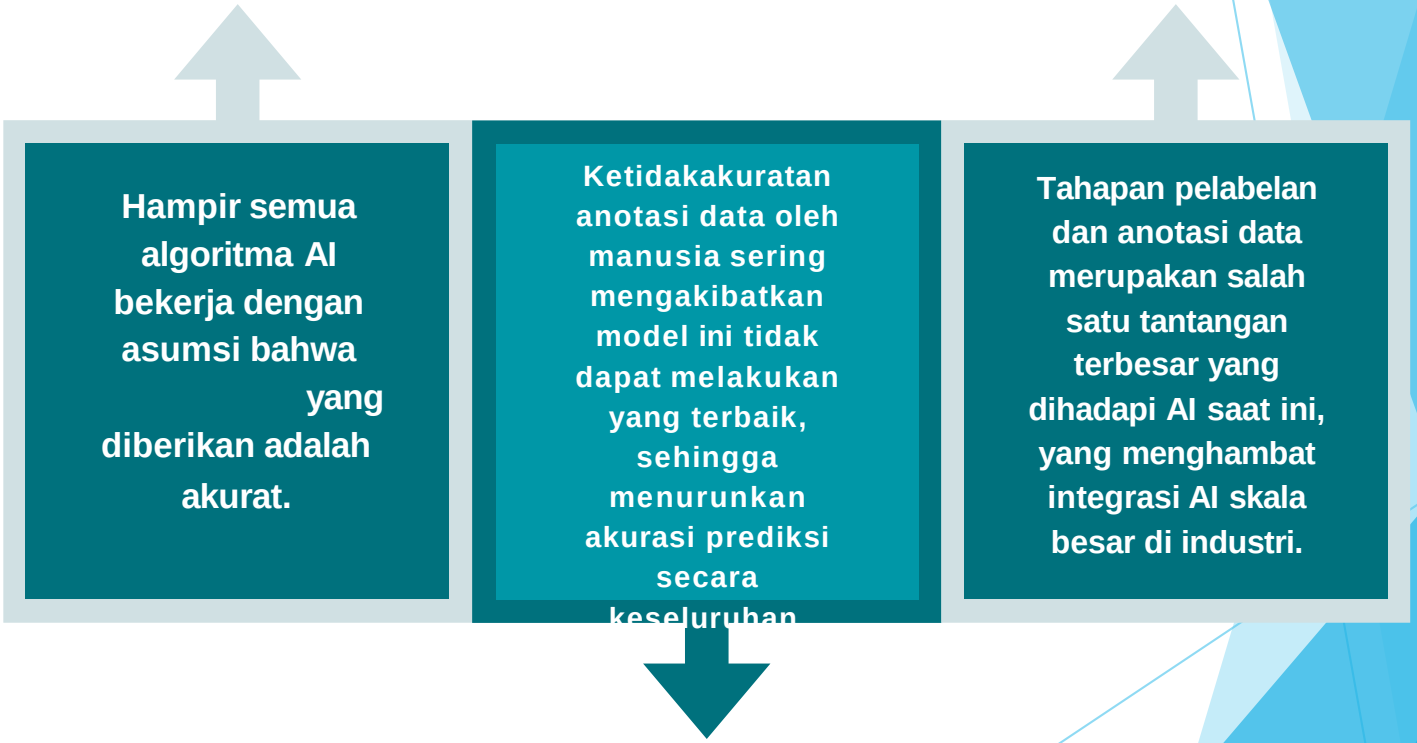
- Phonetic Annotation
  - Anotasi fonetik mengacu pada pelabelan koma dan titik koma yang ada dalam teks dan sangat diperlukan dalam chatbot yang menghasilkan informasi tekstual berdasarkan input yang diberikan kepada mereka.
  - Koma dan berhenti di tempat yang tidak diinginkan dapat mengubah struktur kalimat, menambah pentingnya langkah ini.

Pelabelan Data : Best Practise

- Dengan pembelajaran yang diawasi menjadi bentuk pembelajaran mesin yang paling umum saat ini, pelabelan data ditemukan di hampir setiap tempat kerja yang membahas tentang AI.
- Berikut adalah beberapa praktik terbaik untuk pelabelan data untuk AI guna memastikan model Anda tidak rusak karena data yang buruk:



Pelabelan Data : Best Practise



Dokumentasi Pelabelan Data

Kualitas dan Akurasi Data

- **Akurasi** dalam pelabelan data mengukur **seberapa dekat pelabelan dengan *ground truth*,** atau seberapa baik fitur berlabel dalam data set konsisten dengan kondisi dunia nyata. Misal dalam *computer vision*, dalam meletakkan kotak pembatas di sekitar objek di satu jalanan) atau model pemrosesan bahasa alami (NLP) seperti mengklasifikasikan teks untuk sentimen sosial.

- **Kualitas** dalam pelabelan data adalah tentang a**kurasi dataset secara keseluruhan.** Apakah pekerjaan semua pemberi label terlihat sama? Apakah pelabelan secara konsisten akurat di seluruh data set? Misalkan kita memiliki 29, 89, atau 999 pelabel data yang bekerja secara bersamaan.

Menganalisis Akurasi Pelabelan Data

- Business Goals suatu AI yang berbeda memerlukan ukuran kualitas data yang berbeda.

- Keseimbangan dan variasi titik data di dalam dataset merupakan indikator seberapa baik algoritma dapat memprediksi suatu titik atau pola selanjutnya.

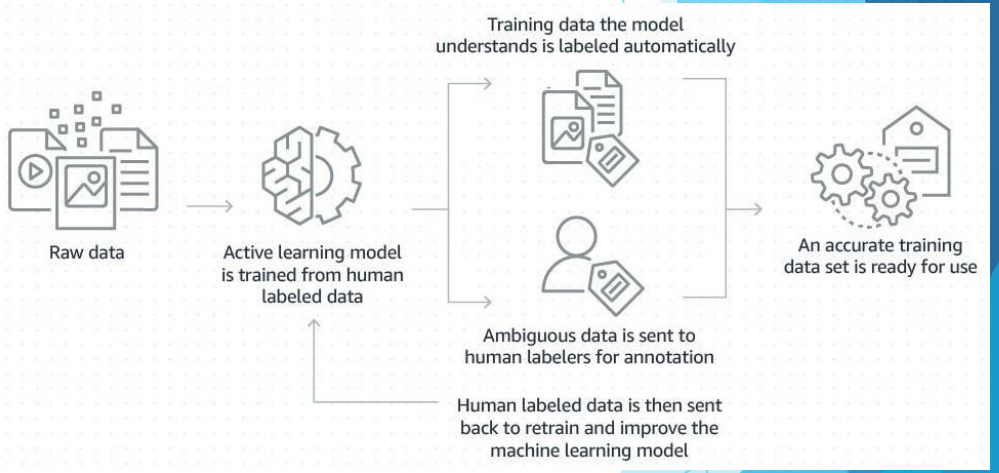
    - Misal tugas suatu AI adalah membedakan antara kendaraan yang bergerak dan tidak bergerak. Jika dataset memuat 90% gambar mobil bergerak tetapi hanya 10% yang diparkir, maka dapat dianggap tidak seimbang.
    - Untuk mengatasi masalah ini dapat digunakan teknik seperti oversampling, downsampling atau weight balancing.

Menganalisis Akurasi Pelabelan Data

- Kualitas data set untuk pelatihan model sering ditentukan oleh seberapa tepat label dan kategori ditempatkan pada setiap titik data.

- Namun, bukan hanya tentang keakuratan pelabelan data tetapi juga tentang seberapa konsisten keakuratannya.

    - Akurasi dan konsistensi data diukur selama proses penjaminan mutu, langkah-langkah terpisah yang dapat dilakukan secara manual atau otomatis.
    - Pendekatan yang berbeda dapat digabungkan untuk *cross check* dan memastikan kesempurnaan data set.



Apa yang mempengaruhi kualitas data dalam pelabelan?

- *Knowledge and context*
  - Pengetahuan dasar satu domain dan pemahaman kontekstual sangat penting seperti pemberi label untuk membuat set data terstruktur berkualitas tinggi.
- *Agility*
  - Pelabelan data berkembang saat dilakukan pengujian dan validasi model, sehingga harus disiapkan data set baru dan memperkaya data set yang ada untuk meningkatkan hasil algoritma Machine Learning.
- *Relationship*
  - Kita memerlukan pemberi label data yang dapat merespons dengan cepat dan mengikuti alur kerja tim, berdasarkan apa yang telah dipelajari dalam fase pengujian dan validasi model.
- *Communication*
  - Pendekatan umpan balik (feedback) adalah cara terbaik untuk membangun komunikasi dan kolaborasi yang andal antara tim dan pemberi label data.

Metode QA untuk Mengukur Kualitas Data

- **Consensus Algorithm**
  - Merupakan proses untuk mencapai reliabilitas data melalui kesepakatan pada satu titik data di antara beberapa individu pemberi label data atau suatu organisasi.
  - Konsensus dapat dilakukan dengan menetapkan sejumlah *reviewer* per titik data
    - (umumnya untuk *data open source*) atau sepenuhnya otomatis.
- **Benchmarking and Gold Standard**
  - Benchmarking adalah pendekatan yang lebih kompleks dan andal untuk QA, karena menggunakan standar tertentu.
  - Menggunakan otomatisasi, pemberi label mendapatkan benchmark secara acak

Metode QA untuk Mengukur Kualitas Data

- **Sample review**
  - Pilih sampel acak dari hasil pelabelan yang telah diselesaikan.
  - Pekerja yang lebih berpengalaman, seperti pemimpin tim atau manajer proyek, dapat meninjau sampel untuk mengukur akurasinya.

- **Cronbach's Alpha Test**
  - Digunakan sebagai ukuran korelasi rata-rata atau konsistensi item dalam dataset, yang tergantung pada karakteristik penelitian (misal homogenitas).
  - Dapat membantu dengan cepat melihat keandalan label secara keseluruhan.

Cronbach’s Alpha Test

- Dikenal sebagai ukuran konsistensi internal yang digunakan dalam konteks instrumen pengukuran multi-item dan memiliki aplikasi yang luas.
- Cronbach's Alpha digunakan untuk mengestimasi item data dalam dataset termasuk label.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right)$$

- Dimana α adalah koefisien reliabilitas, *k* adalah jumlah item set, s_i^2 adalah nilai *variance* setiap item *i* dimana i = *1, 2, …, k,* and s_x^2 adalah nilai *variance* dari semua item. Semakin tinggi nilai koefisien α, maka setiap item memiliki nilai *covariance* dan dapat dihitung (memiliki kesamaan konsep).
- Kategori reliabilitas tinggi dengan nilai α > 0.05.

Cronbach's Alpha Test

- Besarnya koefisien reliabilitas berhubungan langsung dengan skor standar deviasi yang diperoleh dari sampel data apa pun karena koefisien reliabilitas adalah koefisien korelasi.

Series Number	N of Respondents	N of Items	Range	Variance	Standard Deviation	α	Mean
1	200	25	27.00	27.05	5.20	0.10	69.54
2	200	25	30.00	24.36	4.93	0.01	74.29
3	200	25	35.00	42.18	6.49	0.47	80.13
4	200	25	40.00	61.79	7.86	0.67	81.10
5	200	25	45.00	65.50	8.09	0.67	80.42
6	200	25	51.00	68.48	8.27	0.68	79.83
7	200	25	60.00	79.55	8.91	0.74	80.83
8	200	25	72.00	76.62	8.75	0.72	80.93
9	200	25	89.00	103.97	10.19	0.80	80.57
10	200	25	100.00	108.52	10.41	0.81	80.63

- Dapat dilihat bahwa semakin tinggi nilai Range dan Variance membuat nilai reliabilitas juga naik.

Keamanan Pelabelan Data

- What are the security risks of outsourcing data labeling?
  - Mengakses data dari jaringan yang tidak aman atau menggunakan perangkat tanpa perlindungan malware
  - Mengunduh atau simpan sebagian data (mis., screen capture, flash drive)
  - Memberi label data saat berada di tempat umum
  - Tidak memiliki pelatihan, konteks, atau akuntabilitas terkait dengan aturan keamanan untuk pekerjaan labeling
  - Bekerja di lingkungan fisik atau digital yang tidak disertifikasi untuk mematuhi peraturan data (mis., HIPAA, SOC 2).

- Tiga area yang perlu menjadi perhatian untuk menjaga keamanan dokumen
  - **Orang dan Tenaga Kerja:** Ini dapat mencakup pemeriksaan latar belakang untuk pekerja dan mungkin mengharuskan pemberi label untuk menandatangani perjanjian kerahasiaan (NDA) atau dokumen serupa yang menguraikan persyaratan keamanan data.
  - **Teknologi dan Jaringan:** Pekerja mungkin diminta untuk menyerahkan perangkat yang mereka bawa ke tempat kerja, seperti ponsel atau tablet.
  - **Fasilitas dan Ruang Kerja:** Pekerja dapat duduk di tempat yang menghalangi orang lain untuk melihat pekerjaan mereka.

Integrasi Data

Menggabungkan Data

- Dalam beberapa kasus permasalah, dibutuhkan data yang berasal dari beberapa sumber.
- Sumber yang berbeda umumnya memiliki perbedaan pada konvensi penamaan, cara pengelompokan data yang berbeda, dll.
- Sebelum Anda dapat masuk ke bagian penjelajahan dan pembuatan model, Anda harus terlebih dahulu menggabungkan beberapa kumpulan data ini (dalam bentuk tabel, kerangka data, dll.).

Bagaimana Anda bisa melakukan ini tanpa kehilangan informasi?

Menggabungkan Data

Dalam kasus ini, umumnya terdapat dua skenario :

Sesi Lab 10.4

- **Pertama**, data dengan atribut yang sama dapat didistribusikan ke beberapa file.
  - Misalnya, Anda diberikan beberapa file yang masing-masing menyimpan informasi penjualan yang terjadi pada minggu tertentu dalam setahun. Dengan demikian, Anda akan memiliki 52 file sepanjang tahun. Setiap file akan memiliki nomor dan nama kolom yang sama.
- **Kedua**, Anda mungkin perlu menggabungkan informasi dari berbagai sumber.
  - Misalnya, Anda ingin mendapatkan informasi kontak orang-orang yang telah membeli produk Anda. Di sini Anda memiliki dua file – yang pertama berisi informasi penjualan dan yang kedua berisi informasi tentang pelanggan.

Summary

- Dokumentasi juga dilakukan untuk proses transformasi data, seleksi fitur maupun pelabelan data
- Pelabelan bergantung pada pernyataan masalah, kerangka waktu proyek, dan jumlah orang yang terkait dengan pekerjaan

Referensi

- Ozdemir, Sinan Susarla, Divya - Feature engineering made easy identify unique features from your dataset in order to build powerful machine learning systems (2018, Packt Publishing)
- Dong, Guozhu, Liu, H. - Feature Engineering For Machine Learning and Data Analytics
- Jake VanderPlas - Python Data Science Handbook, (2016, O'Reilly Media)
- Soledad Galli - Python Feature Engineering
- Hasib Zunair, Improving performance of image classification models using pretraining and a combination of labeled and unlabeled data, 2020

Tools / Lab Online

- Jupyter Notebook
- Google Collabs

Terima Kasih