

Pengantar Data Sains

#3 Meeting

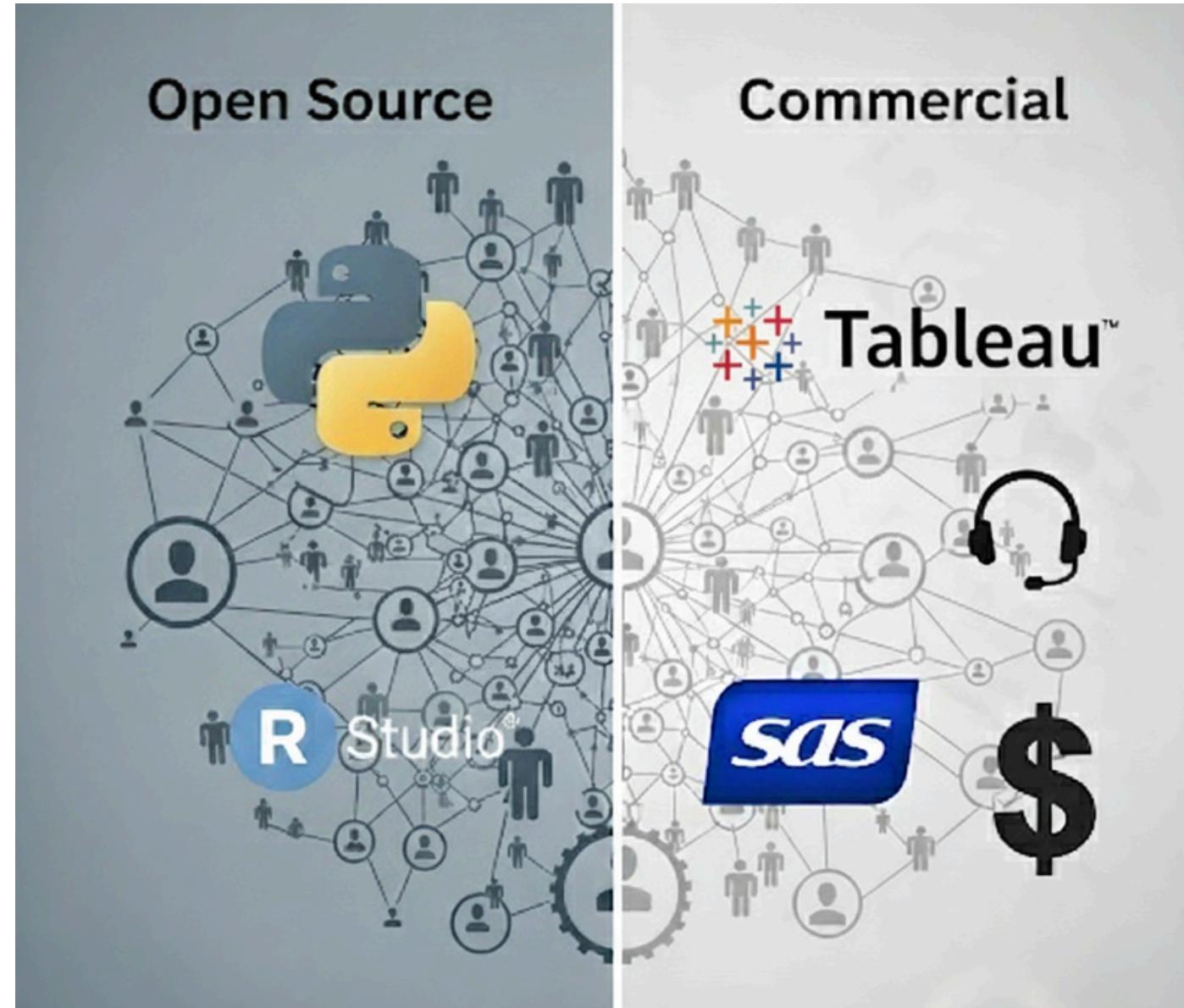
Software dan Analisis

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 199005202024061001



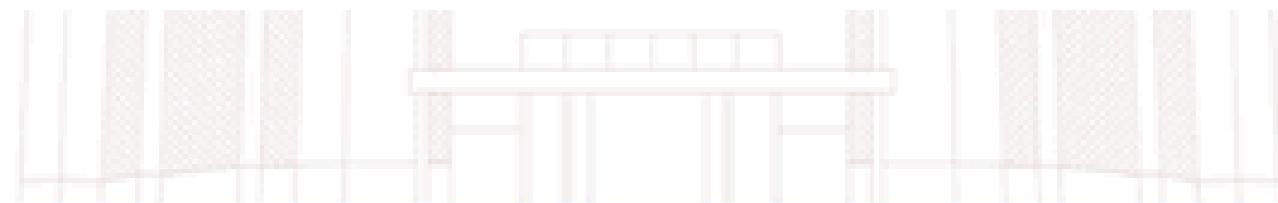


Kategori Software: Model Lisensi



1. Open Source (Sumber Terbuka)
 - Gratis untuk digunakan, dimodifikasi, dan didistribusikan.
 - Didukung oleh komunitas global yang besar dan aktif.
 - Inovasi cepat dan kaya akan library/paket.
 - Contoh: Python, R, SQL (PostgreSQL), Julia, Apache Spark.
 2. Commercial / Proprietary (Berbayar)
 - Membutuhkan lisensi berbayar untuk digunakan.
 - Dilengkapi dengan dukungan pelanggan (customer support) resmi.
 - Antarmuka yang lebih ramah pengguna (berbasis GUI).
 - Contoh: SAS, SPSS, Tableau, Microsoft Power BI, MATLAB.
- Open source memberikan fleksibilitas dan kekuatan tanpa biaya, tetapi kurva belajarnya bisa lebih curam dan tidak ada "customer service" resmi. Software komersial lebih mudah digunakan untuk pemula dan ada dukungan resmi, tetapi biayanya bisa sangat mahal dan kurang fleksibel.

Gartner Magic Quadrant for Analytics and Business Intelligence Platforms





Kategori Software: Lokasi Komputasi



1. On-Premise (Lokal)

- Software diinstal dan dijalankan di komputer atau server lokal milik sendiri.
- Memberikan kontrol penuh atas data dan keamanan.
- Sumber daya terbatas oleh spesifikasi hardware yang dimiliki.
- Contoh: Menginstal Python/RStudio di laptop, server database perusahaan.

2. Cloud (Awan)

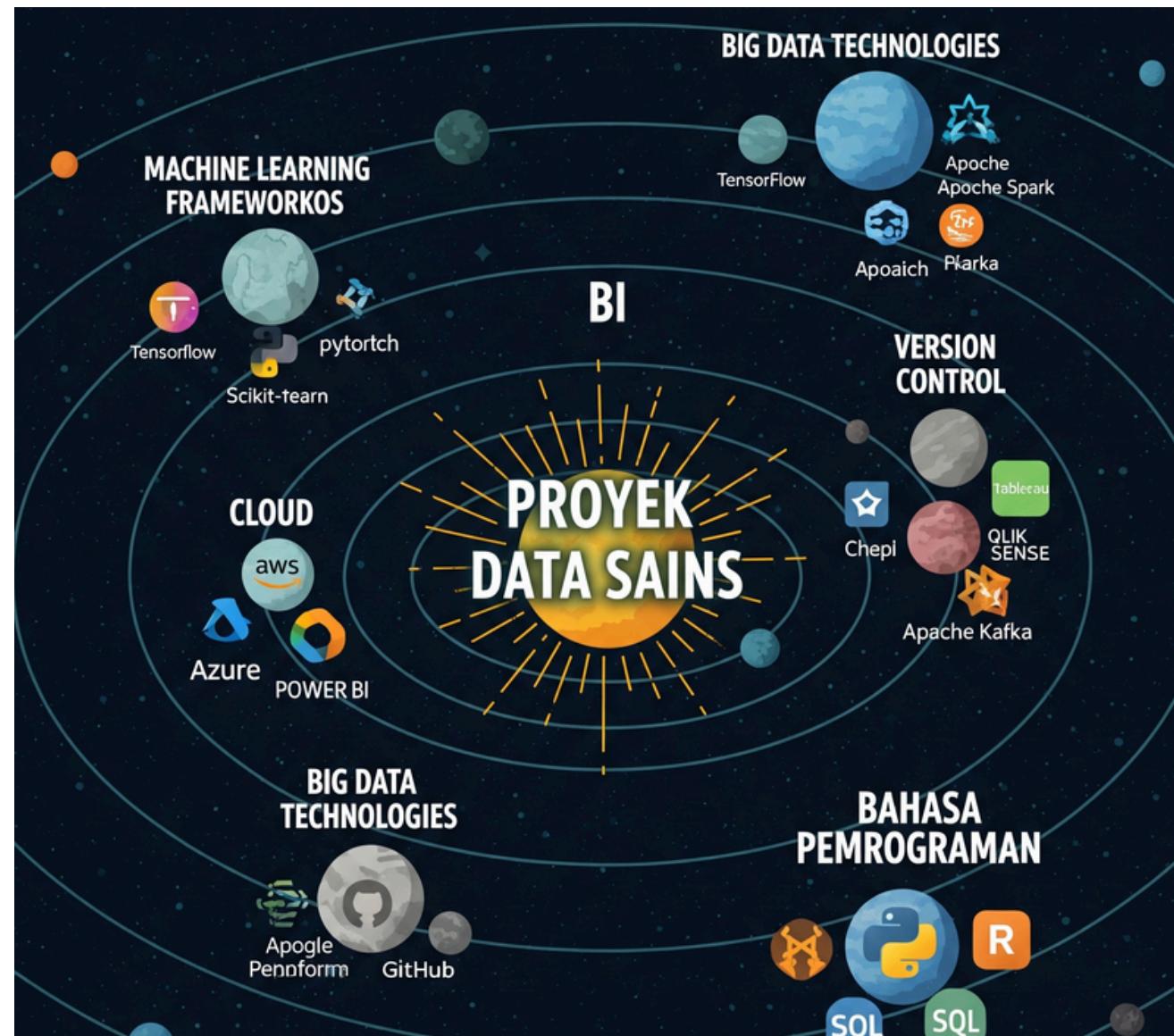
- Software dan infrastruktur diakses melalui internet, dikelola oleh pihak ketiga.
- Skalabilitas "sesuai permintaan" (bayar sesuai pemakaian).
- Memudahkan kolaborasi tim secara geografis.
- Contoh: Google Colab, Amazon SageMaker, Microsoft Azure ML, Databricks.

National Institute of Standards and Technology (NIST) Definition of Cloud Computing

Flexera State of the Cloud Report



Peta Ekosistem Software Data Sains



Kategori	Contoh Tools
Bahasa Pemrograman	Python, R, SQL, Julia, Scala
IDE & Notebooks	Jupyter Notebook, RStudio, VS Code
Business Intelligence (BI)	Tableau, Power BI, Looker, Qlik
Database	PostgreSQL, MySQL, SQL Server, Oracle (SQL), MongoDB (NoSQL)
Big Data	Apache Spark, Apache Hadoop, Apache Kafka
Platform Cloud	AWS, Google Cloud Platform, Microsoft Azure



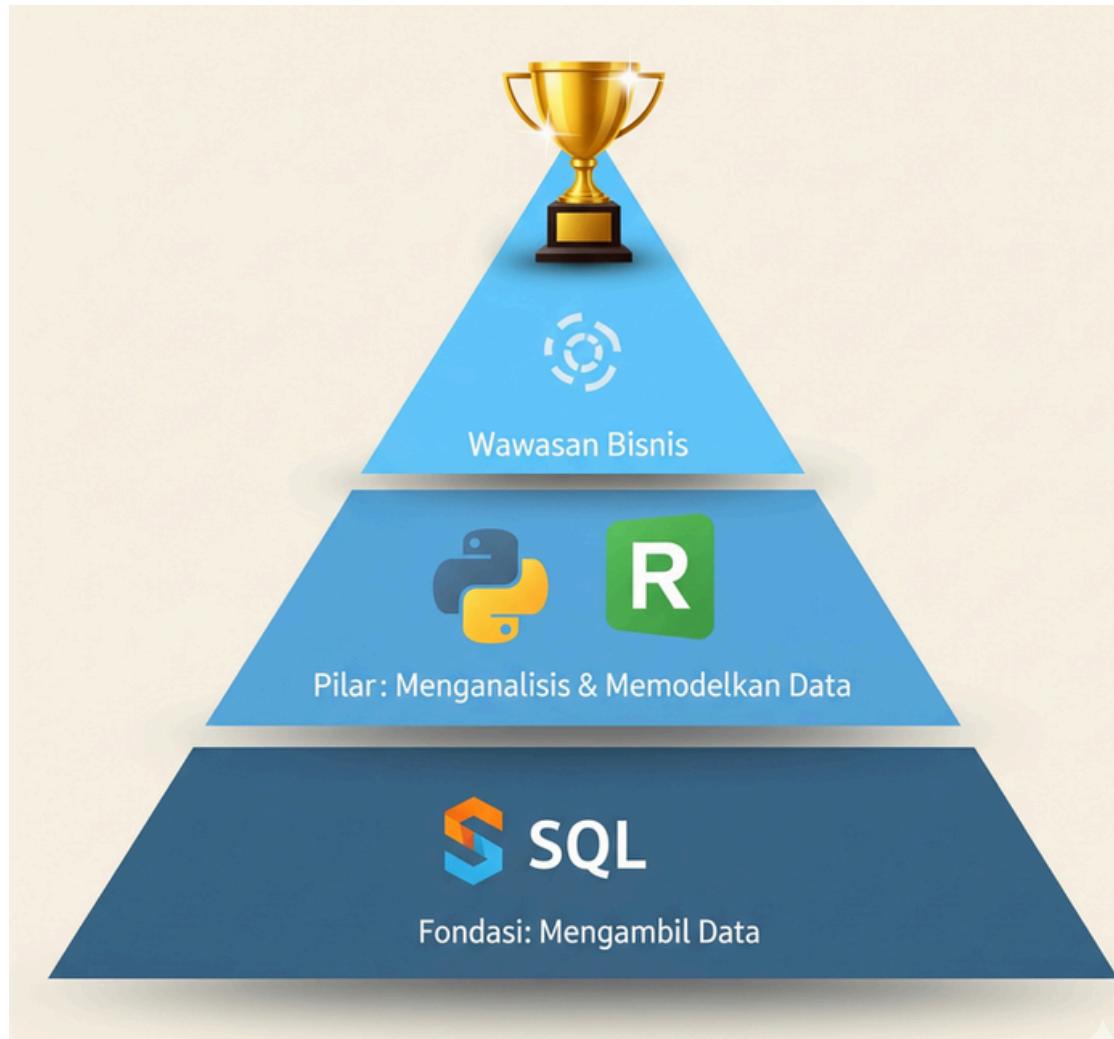
Dunia Bahasa Pemrograman



- Ada ratusan bahasa pemrograman, masing-masing dengan kekuatan dan tujuan yang berbeda.
- Kita bisa mengelompokkannya berdasarkan domain atau tujuan penggunaan utamanya:
- Pengembangan Web (Web Development): Membuat situs web dan aplikasi web interaktif.
 - Contoh: JavaScript, PHP, Ruby, HTML/CSS
- Pengembangan Mobile (Mobile Development): Membuat aplikasi untuk smartphone.
 - Contoh: Swift (untuk Apple iOS), Kotlin (untuk Android)
- Pemrograman Sistem & Game (Systems & Game Programming): Membangun sistem operasi, driver, atau game yang butuh performa tinggi.
 - Contoh: C++, C#, Rust
- Analisis Data & Sains (Data Analysis & Science): Mengekstrak, menganalisis, memodelkan, dan memvisualisasikan data.
 - Contoh: Python, R, SQL, Julia

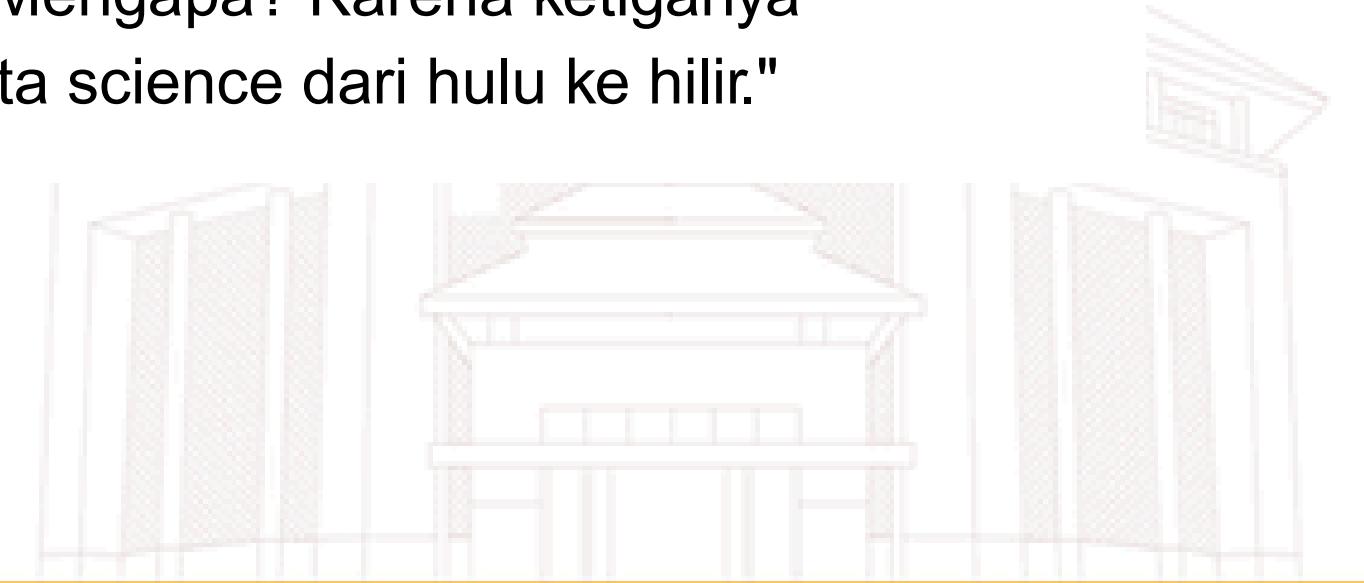


Bahasa Data Sains



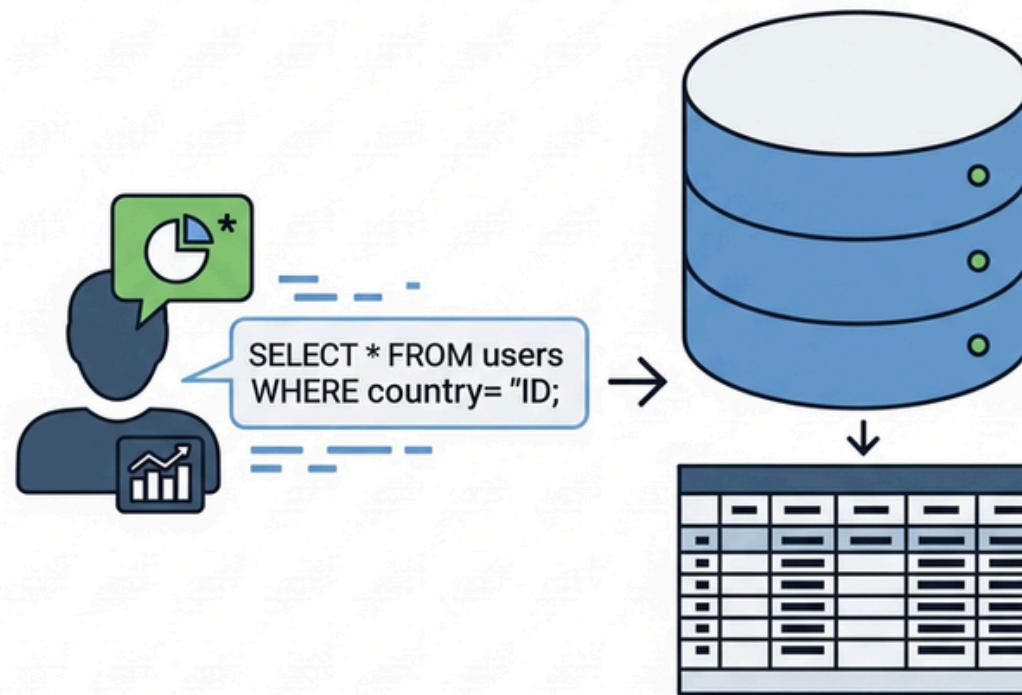
- Di antara puluhan bahasa, tiga nama mendominasi lanskap data: Python, R, dan SQL.
- SQL: Bahasa untuk mengambil dan memanipulasi data di dalam database. Ini adalah fondasi.
- Python & R: Bahasa untuk menganalisis, memodelkan, dan memvisualisasikan data yang sudah diambil.
- Menguasai ketiganya adalah "holy trinity" (tiga serangkai suci) bagi seorang data scientist modern.

"Dari semua bahasa yang ada, kita akan memusatkan perhatian pada tiga alat utama ini. Mengapa? Karena ketiganya membentuk alur kerja data science dari hulu ke hilir."





SQL: Bahasa Universal untuk Data



- SQL (Structured Query Language), diucapkan "sequel".
- Bukan bahasa pemrograman general-purpose, tapi bahasa query.
- Dirancang khusus untuk berinteraksi dengan Database Relasional (data berbentuk tabel).
- Mengapa Wajib?
- Efisiensi: Jauh lebih cepat melakukan filtering, joining, dan agregasi di level database daripada memuat seluruh data ke Python/R.
- Akses Data: Ini adalah satu-satunya cara untuk mendapatkan data dari sebagian besar sistem operasional perusahaan.
- Dialek Populer: PostgreSQL, MySQL, SQL Server, Oracle.

Date, C. J. (2003). An Introduction to Database Systems. Addison-Wesley. (Buku teks klasik yang menjadi fondasi teori database relasional dan SQL).



Python: Sang Generalis

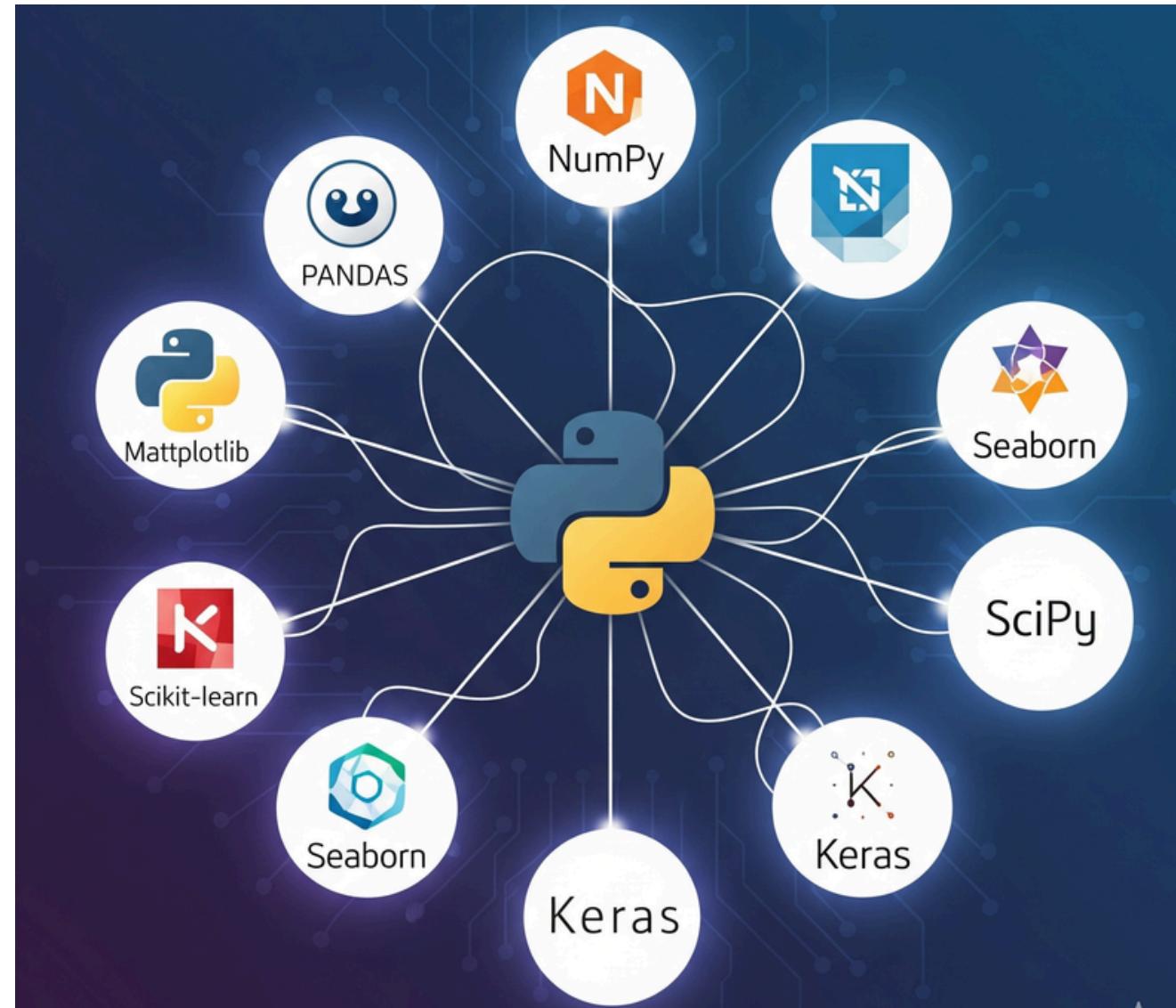


- Dibuat oleh Guido van Rossum pada awal 1990-an.
- Filosofi: Kode yang mudah dibaca dan ditulis (sintaksis bersih).
- Kekuatan dalam Data Sains:
- Serbaguna (General Purpose): Bisa untuk web development, scripting, DAN data science.
- Ekosistem Library yang Luas: Punya "paket" untuk hampir semua kebutuhan.
- Integrasi Mudah: Mudah diintegrasikan dengan sistem lain dan dibawa ke tahap produksi.
- Komunitas Besar: Sangat mudah mencari tutorial dan solusi masalah.
- Dominan di Deep Learning.

Python Software Foundation (PSF)

Van Rossum, G. (1995). Python Tutorial, Technical Report CS-R9526. CWI, Amsterdam.

Ekosistem Python: Paket Esensial



- NumPy: Untuk komputasi numerik. Fondasi untuk semua operasi matematika (Array, Matriks).
- Pandas: Untuk manipulasi dan analisis data terstruktur (DataFrame). Ini adalah "Excel di dalam Python".
- Matplotlib & Seaborn: Untuk visualisasi data. Matplotlib untuk kontrol penuh, Seaborn untuk plot statistik yang indah.
- Scikit-learn: Pustaka machine learning terlengkap (klasifikasi, regresi, clustering, dll).
- TensorFlow & PyTorch: Pustaka utama untuk deep learning (jaringan saraf tiruan).

NumPy adalah "kalkulator super". Pandas adalah "spreadsheet super". Matplotlib/Seaborn adalah "studio seni". Scikit-learn adalah "kotak perkakas machine learning". TensorFlow/PyTorch adalah "pabrik untuk membangun otak buatan".

- McKinney, W. (2017). *Python for Data Analysis*, 2nd Edition. O'Reilly Media.
- Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*.
- Harris, C.R., et al. (2020). "Array programming with NumPy." *Nature*.



R: Sang Spesialis Statistik



- Dibuat oleh Ross Ihaka dan Robert Gentleman pada tahun 1993.
- Dibuat oleh ahli statistik, untuk ahli statistik.
- Kekuatan dalam Data Sains:
- Dibangun untuk Statistik: Fungsi dan struktur datanya dirancang dari awal untuk analisis statistik.
- Visualisasi Data Superior (ggplot2): Dianggap sebagai salah satu library visualisasi data terbaik.
- Komunitas Akademik & Riset yang Kuat: Banyak paket statistik canggih dirilis di R terlebih dahulu.
- RStudio IDE: Lingkungan pengembangan yang sangat matang dan nyaman untuk analisis data.

- The R Foundation for Statistical Computing
- Ihaka, R., & Gentleman, R. (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*.



Ekosistem R: Paket Esensial



- Tidyverse: Sebuah "ekosistem di dalam ekosistem" untuk data science yang modern di R. Termasuk di dalamnya:
- dplyr: Untuk manipulasi data (filter, select, mutate). Sangat intuitif.
- ggplot2: Untuk visualisasi data deklaratif yang sangat kuat.
- tidyverse: Untuk merapikan data (tidy data).
- readr: Untuk membaca data file (seperti CSV) dengan cepat.
- caret / tidymodels: Kerangka kerja komprehensif untuk machine learning.
- Shiny: Untuk membangun aplikasi web interaktif dan dasbor langsung dari R.

Tidyverse, yang dikembangkan oleh Hadley Wickham, merevolusi cara orang bekerja dengan data di R. Sintaksisnya sangat konsisten dan mudah dibaca. Shiny adalah kekuatan super R, memungkinkan analis untuk membuat dasbor interaktif tanpa perlu belajar web development.

- Wickham, H. (2014). "Tidy Data." *The Journal of Statistical Software*.
- Wickham, H., & Grolemund, G. (2016). *R for Data Science*. O'Reilly Media.



Python vs R: Perbandingan Sintaks

R (dengan dplyr)	Python (dengan pandas)
<pre>r sales %>% filter(produk == "Laptop") %>% select(tanggal, pendapatan)</pre>	<pre>python sales[sales['produk'] == 'Laptop'][['tanggal', 'pendapatan']]</pre>
Filosofi: Seperti "kalimat" yang mengalir, menggunakan operator %>% (pipe).	Filosofi: Seperti "mengiris" dan "memilih" data dari sebuah struktur data.

Tugas: Memfilter data penjualan untuk produk "Laptop" dan memilih kolom tanggal & pendapatan.

Sintaksis Tidyverse di R sering dianggap lebih mudah dibaca oleh manusia karena alurnya seperti kalimat. Sintaksis Pandas di Python sangat kuat dan sejalan dengan cara kerja programmer pada umumnya.





Kapan Menggunakan yang Mana?



Gunakan Python jika...

Proyek perlu diintegrasikan dengan aplikasi lain (Produksi).

Anda bekerja di bidang Deep Learning (TensorFlow/PyTorch).

Tim Anda memiliki latar belakang Software Engineering.

Anda butuh satu bahasa serbaguna untuk semua tugas.

Gunakan R jika...

Fokus utama Anda adalah analisis statistik mendalam (Riset).

Anda ingin membuat visualisasi data kompleks untuk laporan (ggplot2).

Tim Anda memiliki latar belakang Statistik/Akademik.

Anda butuh paket statistik canggih yang baru dirilis di dunia riset.



Platform Cloud Terintegrasi



- Penyedia cloud utama menawarkan platform data science end-to-end.
- Menggabungkan penyimpanan data, komputasi, notebooks, machine learning, dan deployment dalam satu atap.
- Tiga Pemain Utama:
 - AWS (Amazon Web Services): Menawarkan Amazon SageMaker.
 - Google Cloud Platform (GCP): Menawarkan Vertex AI.
 - Microsoft Azure: Menawarkan Azure Machine Learning.
- Ini adalah pilihan populer untuk perusahaan karena kemudahan manajemen dan skalabilitas.

"Bayangkan sebuah 'dapur super' yang sudah menyediakan semua bahan, semua peralatan canggih, oven, dan bahkan layanan pengantaran, semua dalam satu tempat. Itulah yang ditawarkan platform cloud ini."

Forrester Wave: AI/ML Platforms: Laporan evaluasi berkala yang membandingkan platform cloud terkemuka (AWS SageMaker, Vertex AI, Azure ML) dan menjadi acuan bagi perusahaan dalam memilih platform.



Laboratorium di Cloud

The screenshot shows the Google Colab interface. At the top, it says "Google Colab". Below that, there are two code cells. The first cell is in Python and contains:

```
fudi from Python"
{
    print("Hello from Python!")

}
print("from Python!")
```

The second cell is in R and contains:

```
fterc..R {
{
    print("Hello Python!")

}
print("Hello from R!")
}
```

Apa itu Google Colaboratory (Colab)?

Sebuah platform Jupyter Notebook gratis yang berjalan di cloud Google, yang secara default mendukung Python namun juga bisa dikonfigurasi untuk bahasa lain.

Mengapa Sangat Populer untuk Pemula & Profesional?

- Tanpa Instalasi (Zero-Setup): Lingkungan Python siap pakai dalam hitungan detik. Cukup buka browser.
- Akses Hardware Gratis: Menyediakan akses gratis ke GPU dan TPU, sangat penting untuk komputasi berat seperti deep learning.
- Kolaborasi Mudah: Bekerja seperti Google Docs, notebook bisa dibagikan dan diedit bersama.
- Dukungan Multi-Bahasa: Meskipun identik dengan Python, Colab juga dapat menjalankan R dan Swift, memberikan fleksibilitas untuk berbagai jenis analisis.

Google Colab FAQ & Seedbank

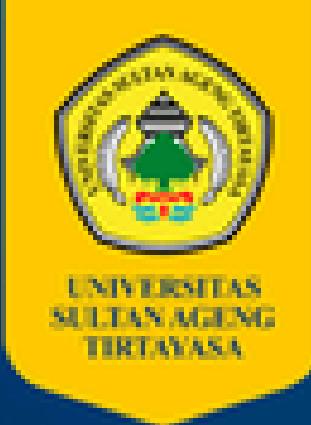
Carneiro, T., et al. (2018). Jupyter Notebooks in the Cloud. (Paper yang relevan dengan arsitektur dan tantangan platform notebook cloud seperti Colab).

Alat Lain yang Perlu Diketahui



- Julia: Bahasa modern yang mencoba menggabungkan kemudahan Python dengan kecepatan C. Populer di komputasi ilmiah dan finansial.
- Scala: Berjalan di atas Java Virtual Machine (JVM). Bahasa pilihan untuk bekerja dengan Apache Spark, framework untuk pemrosesan data skala besar (big data).
- Tableau / Power BI: Bukan bahasa pemrograman, tapi alat Visualisasi & Business Intelligence (BI) berbasis drag-and-drop. Sangat penting untuk komunikasi hasil analisis kepada audiens non-teknis.

Seiring kemajuan karier, seorang data science mungkin perlu mempelajari alat-alat ini. Julia untuk performa tinggi, Scala untuk big data, dan Tableau/Power BI untuk penyajian hasil.



SEE YOU NEXT WEEK !

Ferdian Bangkit Wijaya, S.Stat., M.Si

NIP. 199005202024061001

ferdian.bangkit@untirta.ac.id

