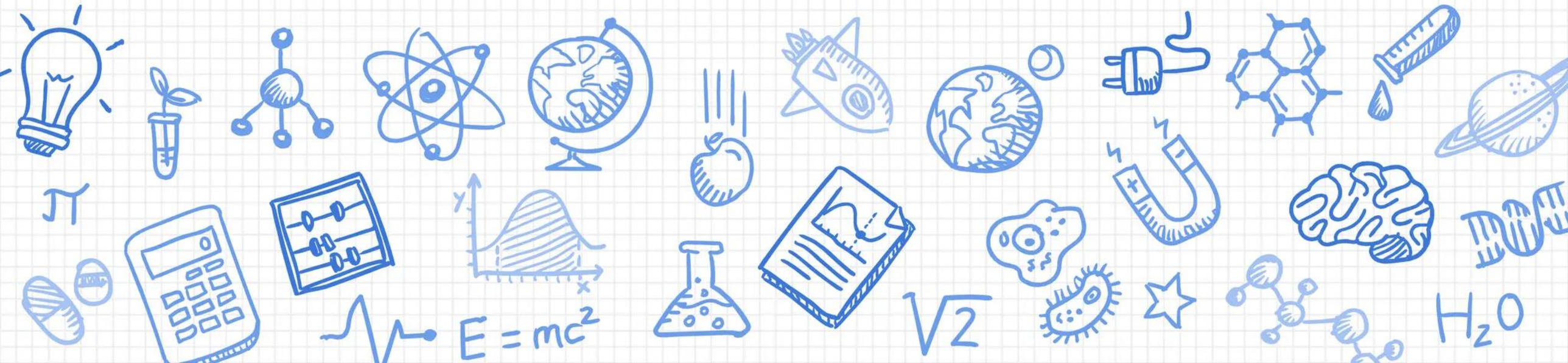
– Data –
Science and Engineering



HELLO!

I am Hendri Karisma

Principal R&D and Data Lead at Akar Inti Data

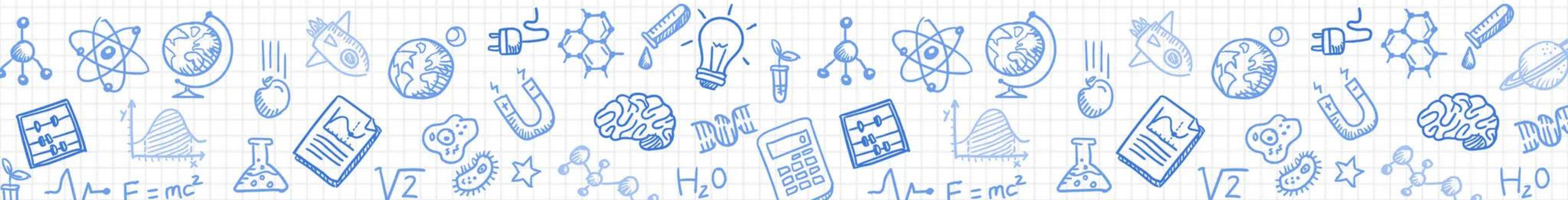
- ✘ Research and Development for Advanced Data Analytics Platform.
- ✘ Data marketplace at nusadata.ai .

You can find me at :

- ✘ telegram @siganteng
- ✘ Instagram @karism4_
- ✘ X @infoHendri

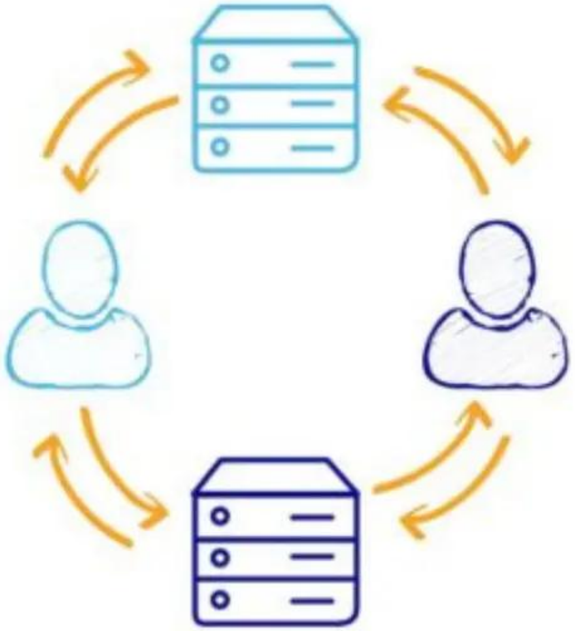
Data

What is data ?

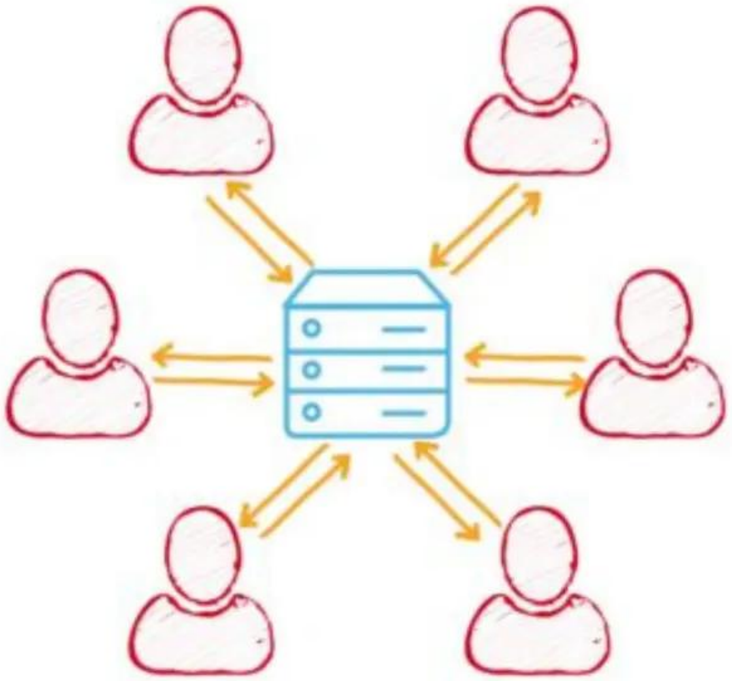


Data are Facts

first-party data



second-party data



third-party data

Data marketplace

Consumer

Single platform

Market Place

Provider

Open Finance

Telco

eKYC

3 Important Statistics About How Much Data Is Created Every Day

FinancesOnline
REVIEWS FOR BUSINESS

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista

3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION
searches on Google by the end of 2021

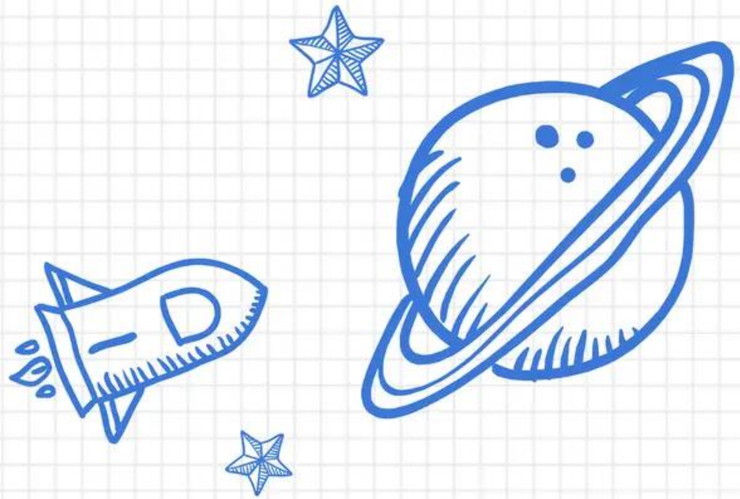
1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

278,108 PETABYTES
global IP data per month by the end of 2021

230,000
new malware versions created every day

82%
share of video in total global internet traffic at the end of 2021



Science vs Engineering

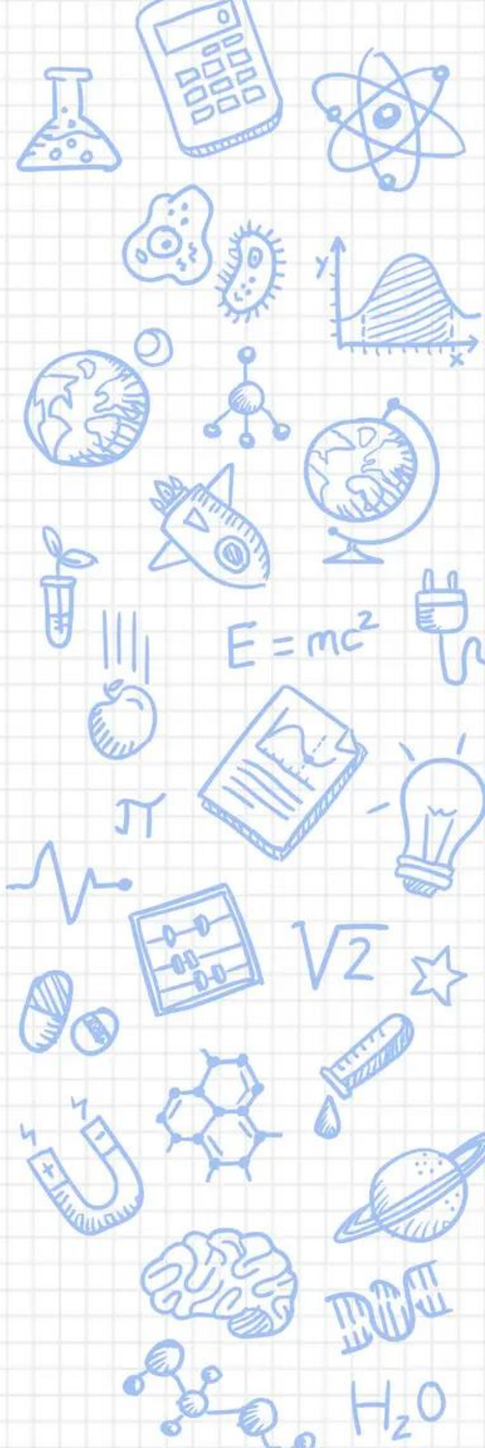
What is the difference?

Science

The goal of a scientist is to answer questions and discover information about their chosen field of study.

Engineering

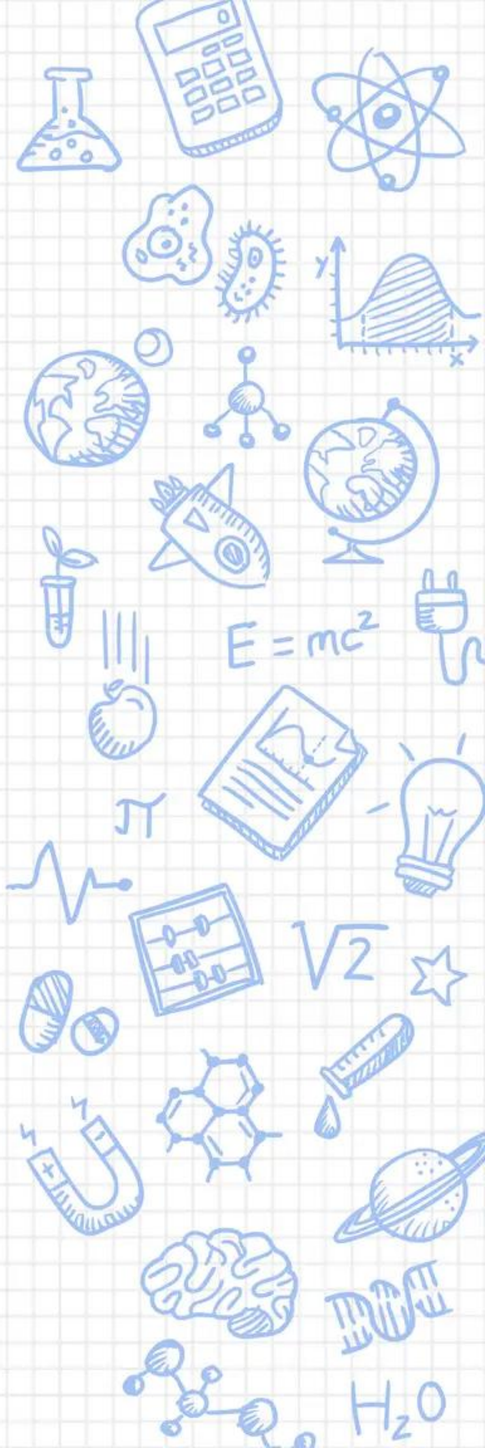
An engineer might produce physical item or blueprint for a new process.



Data roles

- ✘ Data **Scientist**
- ✘ Data Analyst
- ✘ Business Intelligence
- ✘ Data **Engineer**
- ✘ AI/ML **Engineer**

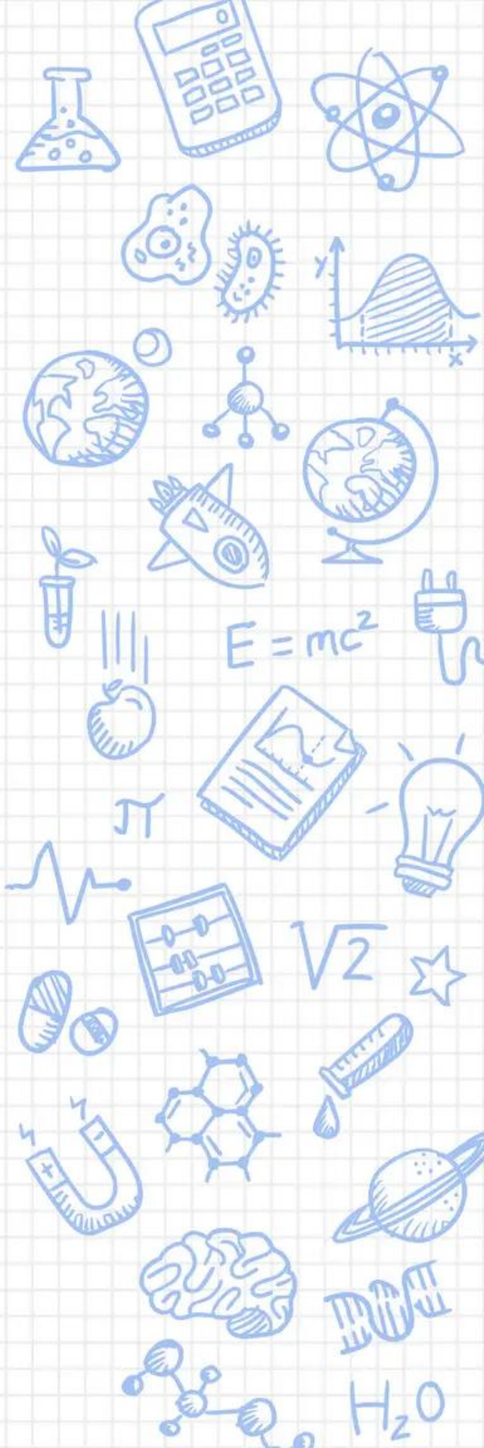
There is a Scientist role and 2 engineer roles.



Data Scientist

The term “data scientist” was coined as recently as 2008 when companies realized the need for data professionals who are skilled in organizing and analyzing massive amounts of data.1 In a 2009 McKinsey&Company article, Hal Varian, Google’s chief economist and UC Berkeley professor of information sciences, business, and economics, predicted the importance of adapting to technology’s influence and reconfiguration of different industries.

Data-driven individuals with high-level technical skills who are **capable of building complex quantitative algorithms** to organize and synthesize large amounts of information used to answer questions and drive strategy in their organization.

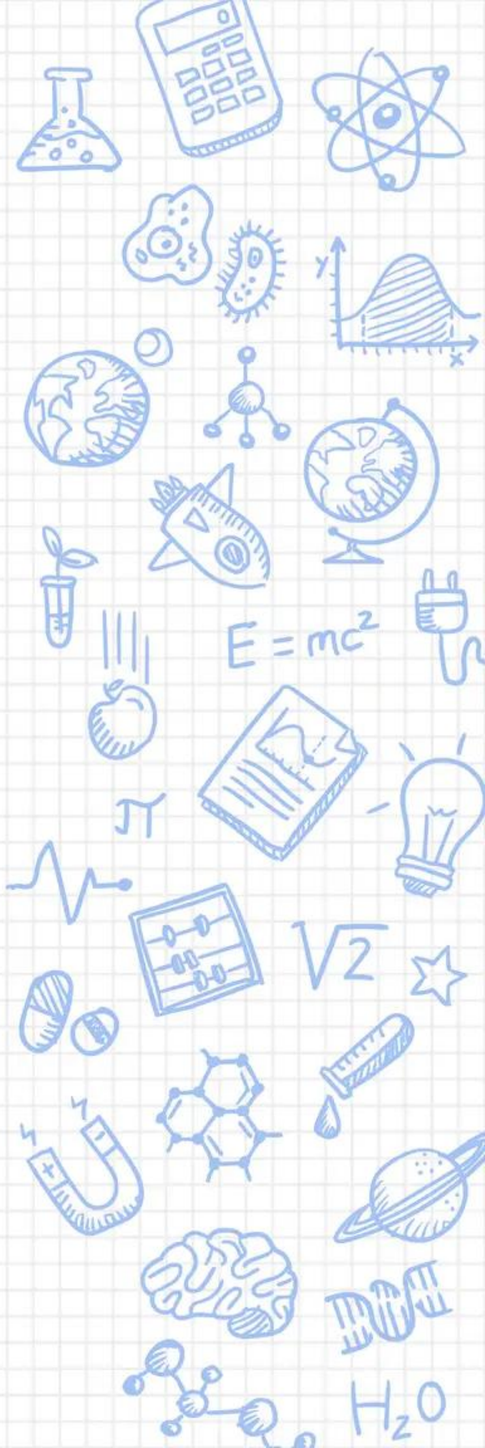


Data Engineer

Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret. Their ultimate goal is to make data accessible so that organizations can use it to evaluate and optimize their performance.

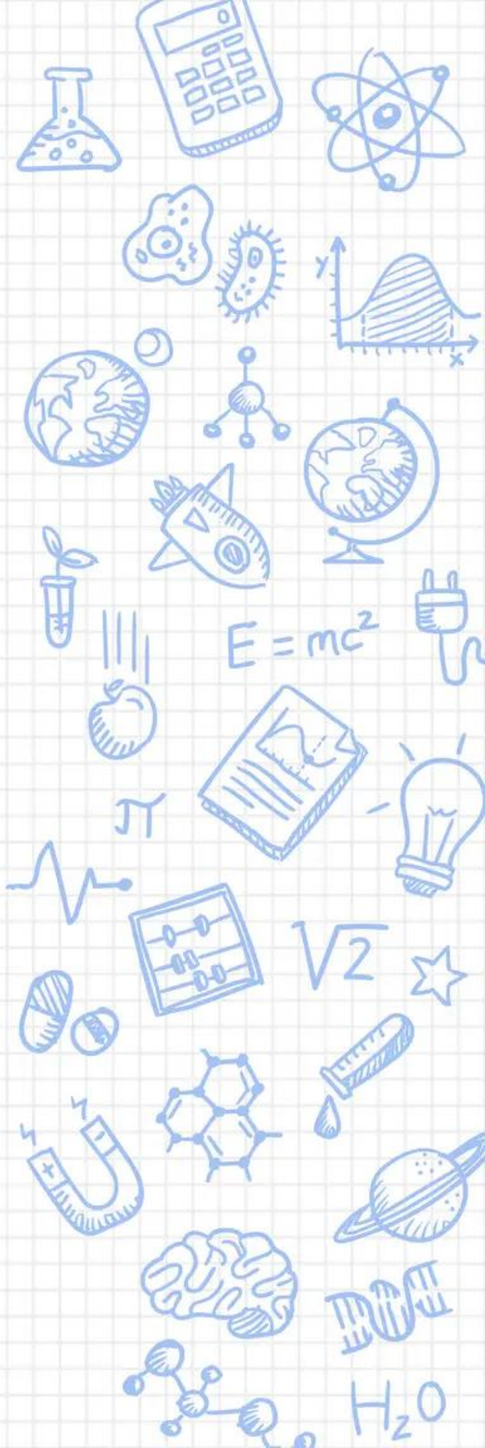
These are some common tasks you might perform when working with data:

- ✘ Acquire datasets that align with business needs
- ✘ Develop algorithms to transform data into useful, actionable information
- ✘ Build, test, and maintain database pipeline architectures
- ✘ Collaborate with management to understand company objectives
- ✘ Create new data validation methods and data analysis tools
- ✘ Ensure compliance with data governance and security policies



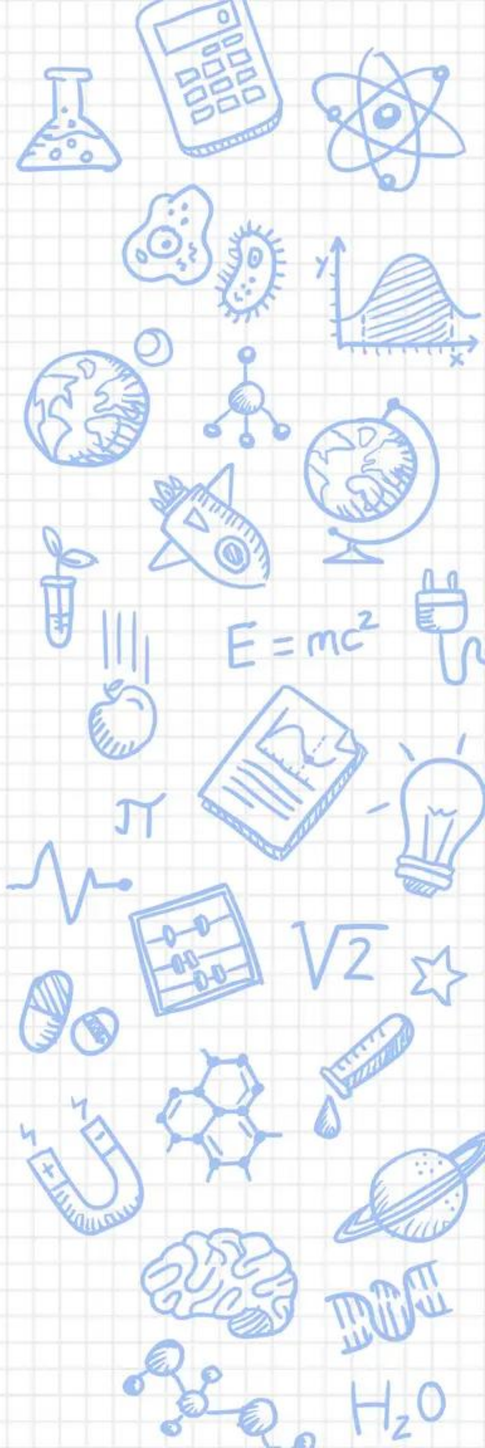
AI Engineer

Artificial intelligence (AI) engineers are responsible for developing, programming and training the complex networks of algorithms that make up AI so that they can function like a human brain. This role requires combined expertise in software development, programming, data science and data engineering. Though this roleis related to data engineering, AI engineers are rarely required to write the code that develops scalable data sharing.



AI Engineer Responsibilities

- ✘ Build AI models from the ground up and explain results to product managers and stakeholders
- ✘ Develop, test, and deploy AI models
- ✘ Convert machine learning models into APIs so other applications can utilize it
- ✘ Build data ingestion and data transformation infrastructure
- ✘ Work alongside data and business analysts
- ✘ Execute statistical analysis and tune results to extract better insights
- ✘ Automate infrastructure used by the data science team
- ✘ Create and manage AI development and production infrastructure



Data Science Experiments

- ✘ Do the data understanding
- ✘ Data exploration
- ✘ Feature Engineering
- ✘ Method/Model Experiments

jupyter spectrogram (autosaved)

File Edit View Insert Cell Kernel Help | Python 3

Markdown CellToolbar

Simple spectral analysis

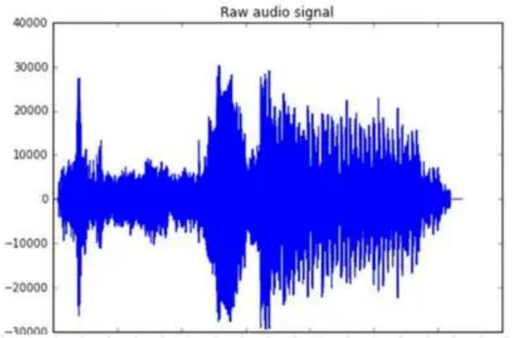
An illustration of the Discrete Fourier Transform

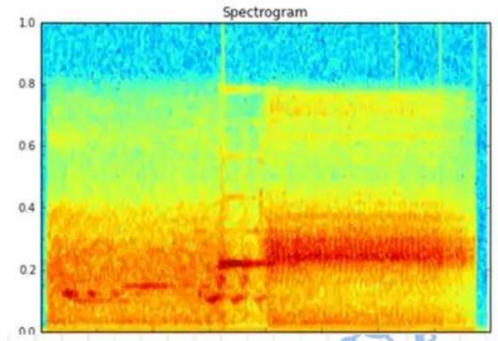
$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N} kn} \quad k = 0, \ldots, N-1$$

```
In [2]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

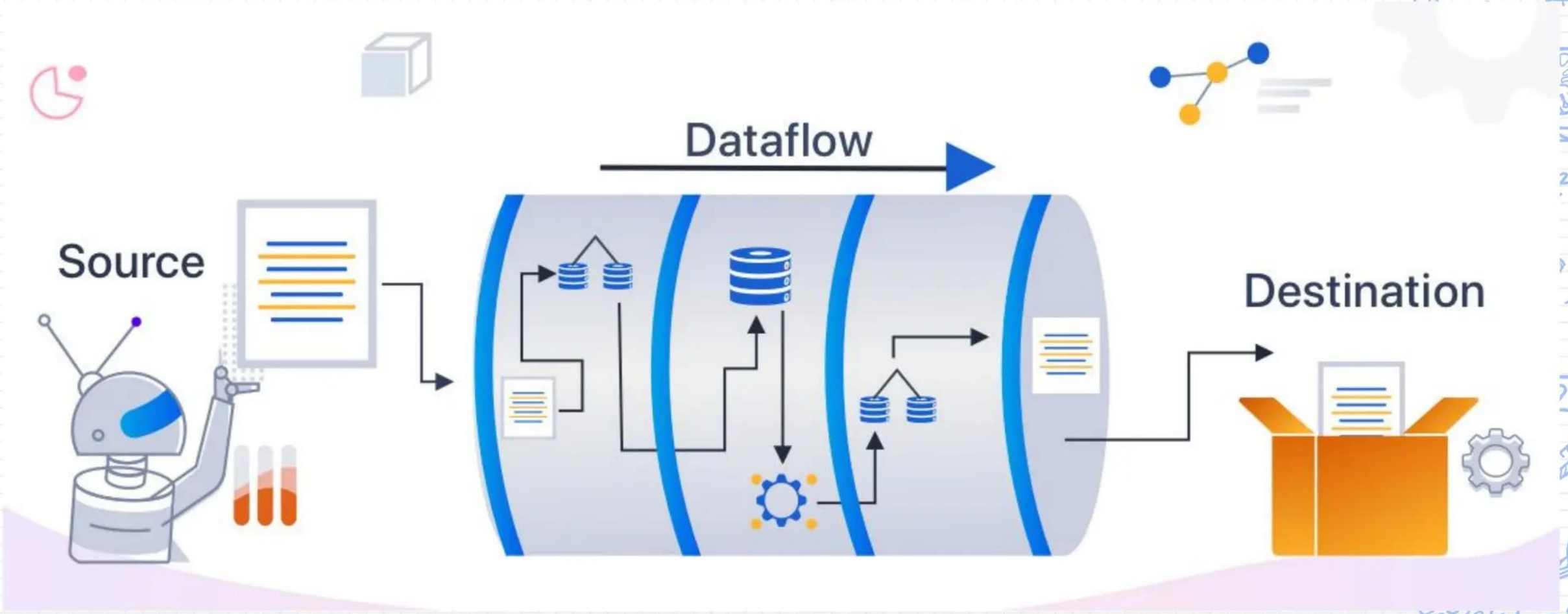
And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

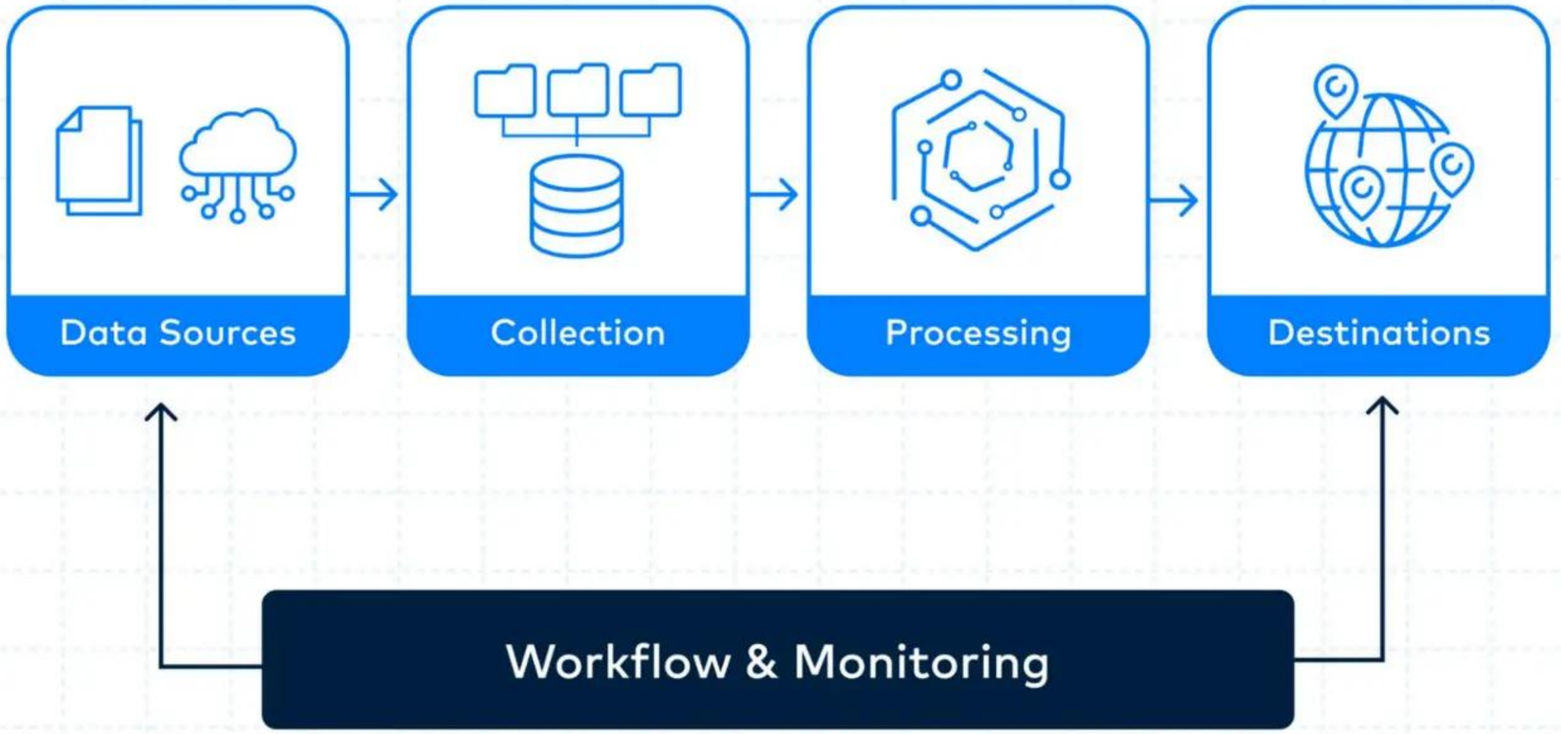




Data flow

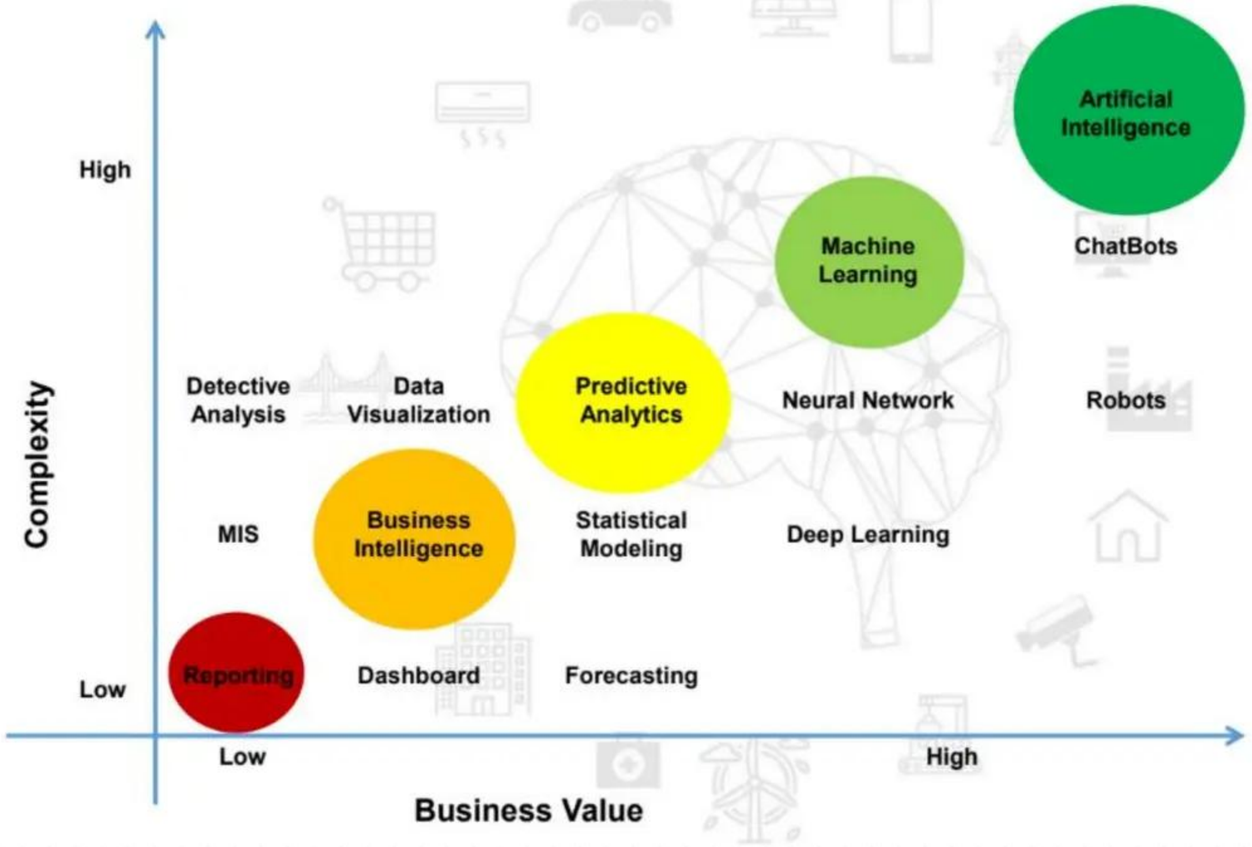


Zoom in


Data Sources
Collection
Processing
Destinations
Workflow & Monitoring

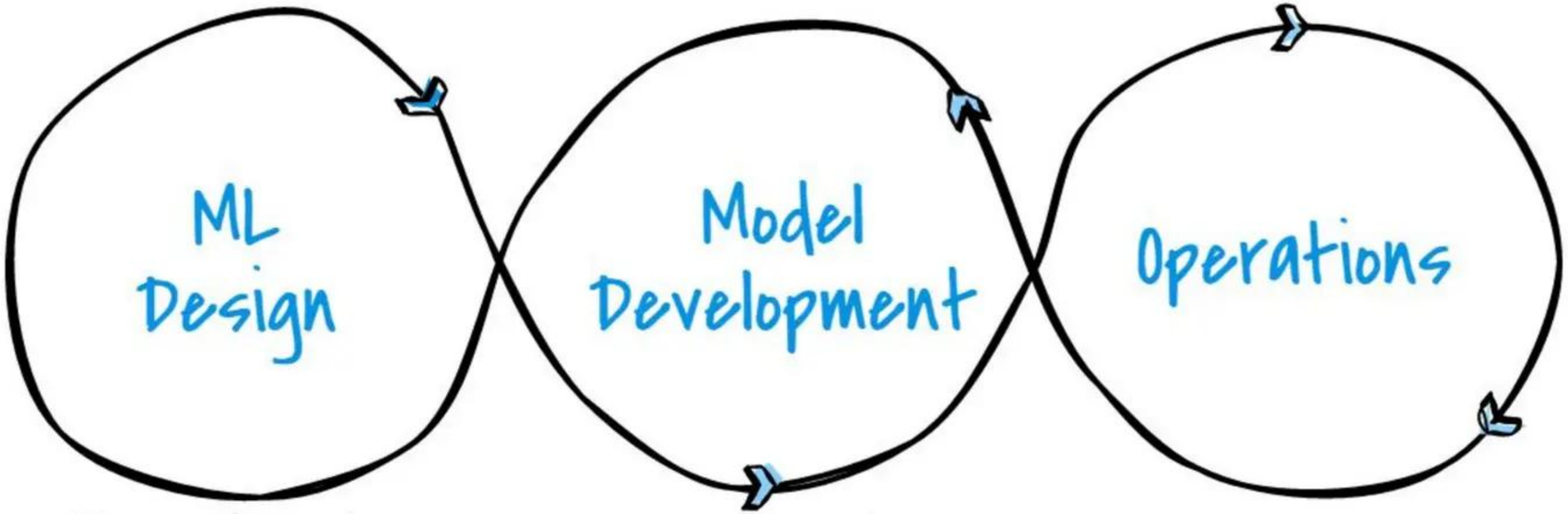
Spectrum of Data Science

Analytics Vidhya
Learn everything about analytics



ML Ops

Machine Learning Operations (MLOps)

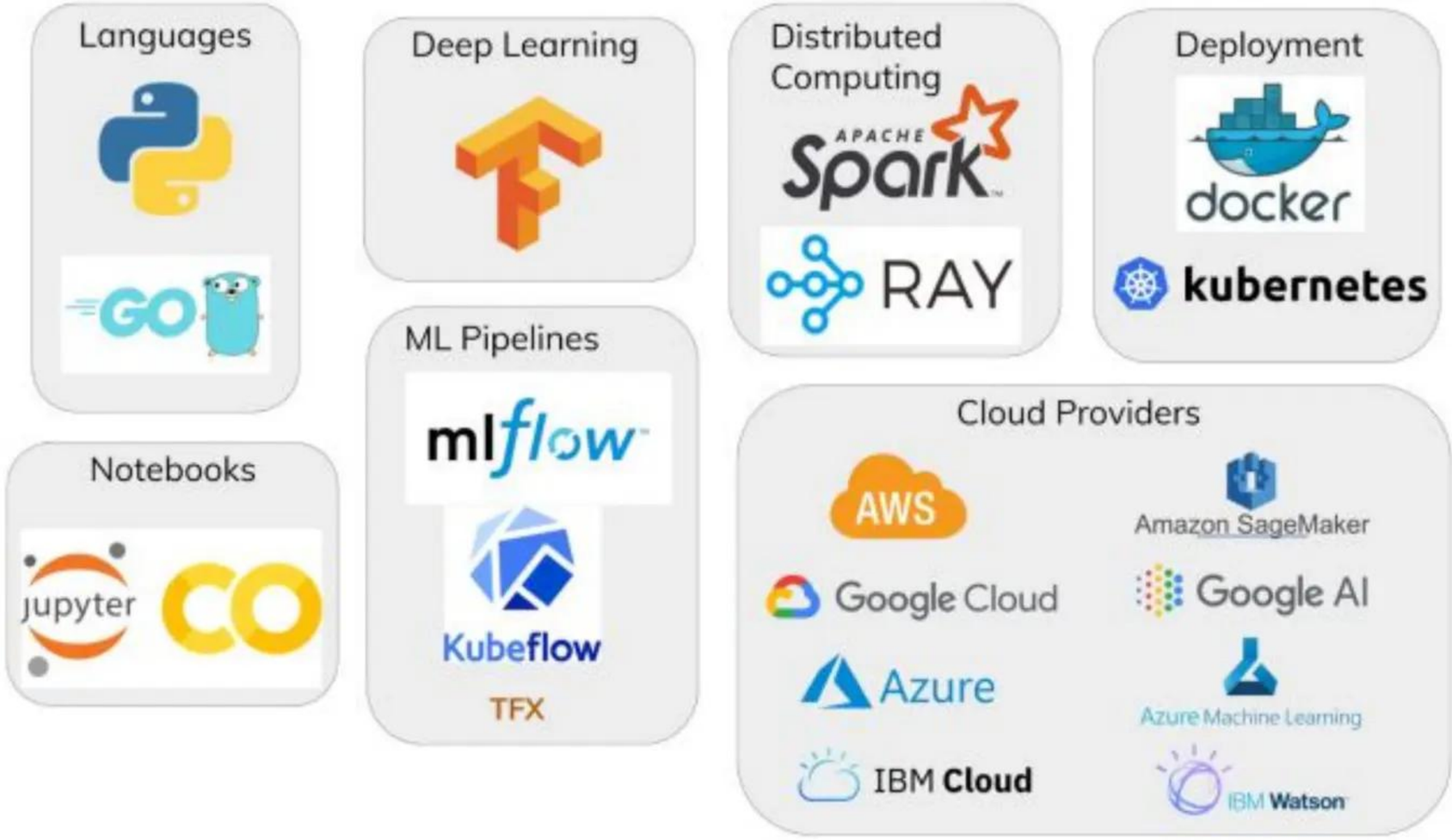


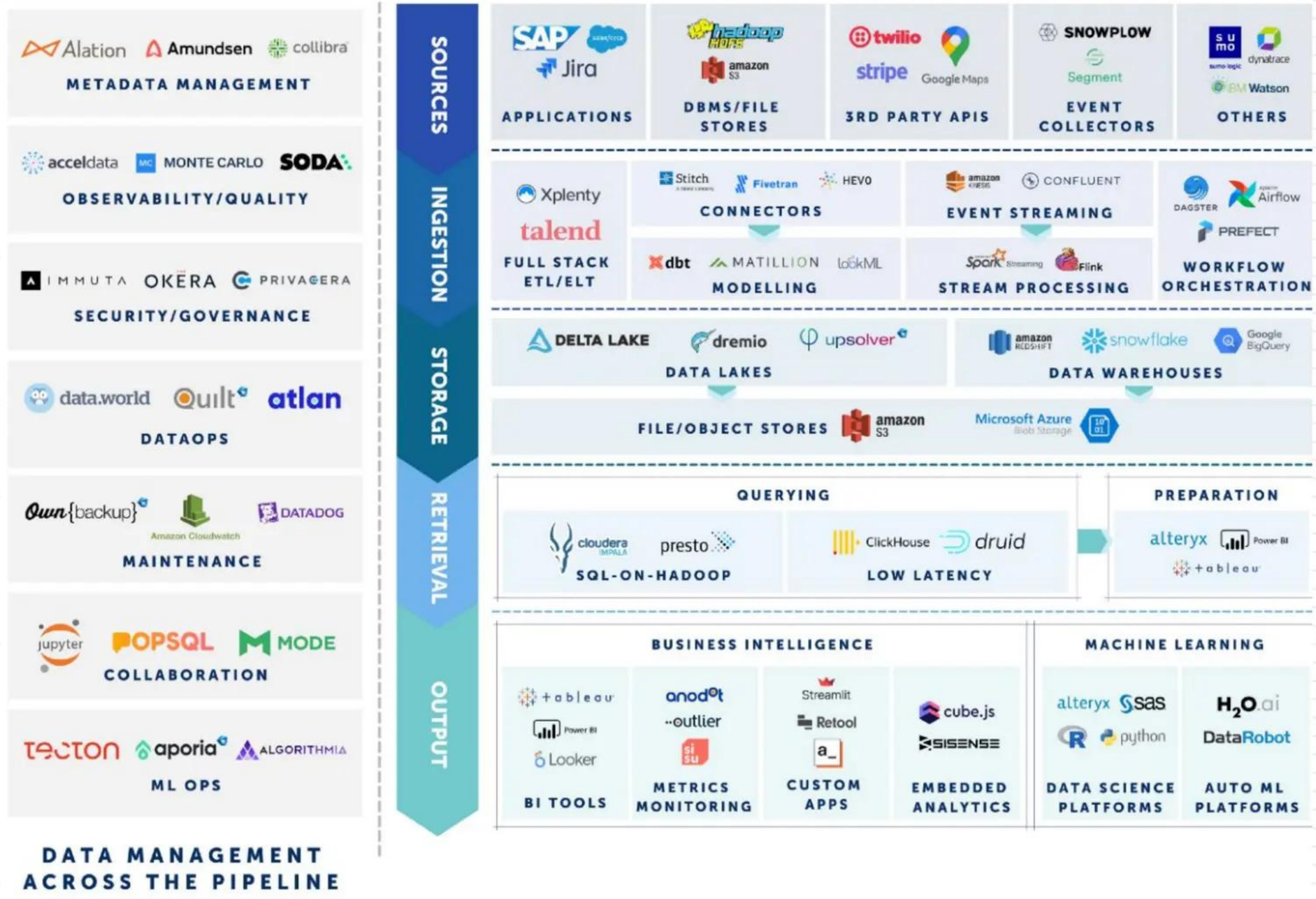
- Gather requirements
- Prioritize ML use cases
- Business understanding
- Data Acquisition

- Data prep & processing
- Feature Engineering
- Model training / experimentation
- Model analysis & evaluation

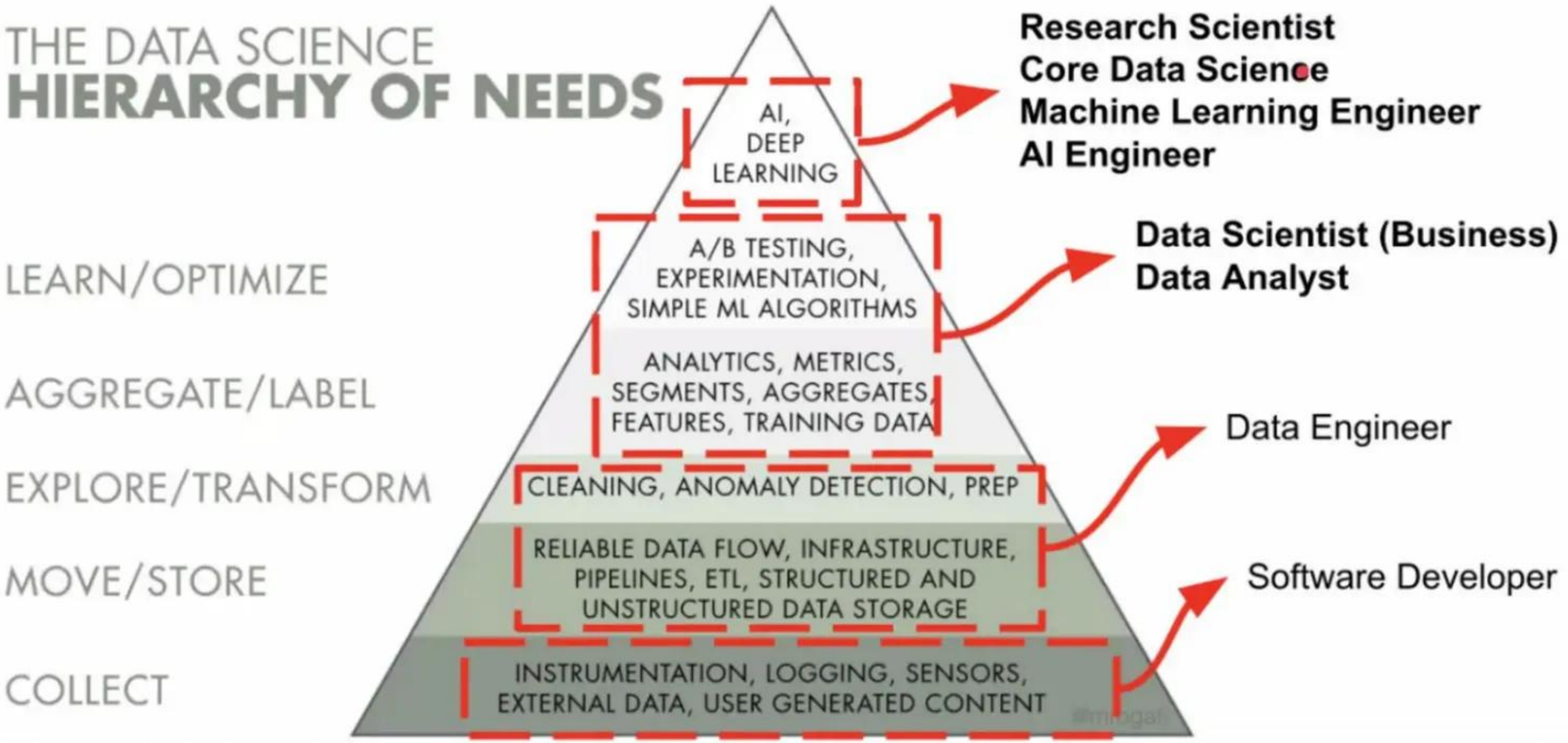
- ML Model Deployment
- CI/CD Pipelines
- Model Monitoring & Triggering

Tech Stack


Languages
Deep Learning
Distributed Computing
Deployment
Notebooks
ML Pipelines
Cloud Providers
Kubeflow
TFX
Amazon SageMaker
Google Cloud
Google AI
Azure
Azure Machine Learning
IBM Cloud
IBM Watson

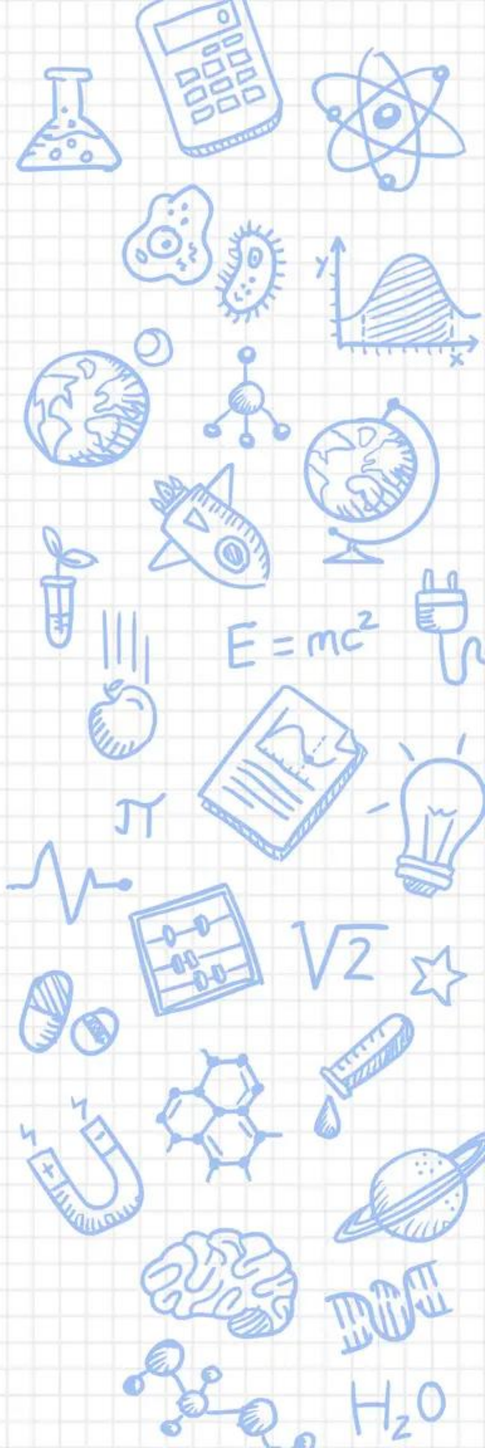

Alation
Amundsen
collibra
METADATA MANAGEMENT
acceldata
MONTE CARLO
SODA
OBSERVABILITY/QUALITY
IMMUTA
OKERA
PRIVACERA
SECURITY/GOVERNANCE
data.world
Quilt
atlan
DATAOPS
Own{backup}
Amazon Cloudwatch
DATADOG
MAINTENANCE
jupyter
POPSQL
MODE
COLLABORATION
tecton
aporia
ALGORITHMIA
ML OPS
DATA MANAGEMENT ACROSS THE PIPELINE
SOURCES
INGESTION
STORAGE
RETRIEVAL
OUTPUT
SAP
salesforce
Jira
APPLICATIONS
hadoop
HDFS
amazon S3
DBMS/FILE STORES
twilio
stripe
Google Maps
3RD PARTY APIS
SNOWPLOW
Segment
EVENT COLLECTORS
sumo logic
dynatrace
IBM Watson
OTHERS
Xplenty
talend
FULL STACK ETL/ELT
Stitch
Fivetran
HEVO
CONNECTORS
dbt
MATILLION
lookML
MODELLING
amazon KINESIS
CONFLUENT
EVENT STREAMING
Spark Streaming
Flink
STREAM PROCESSING
DAGSTER
Airflow
PREFECT
WORKFLOW ORCHESTRATION
DELTA LAKE
dremio
upsolver
DATA LAKES
amazon REDSHIFT
snowflake
Google BigQuery
DATA WAREHOUSES
FILE/OBJECT STORES
amazon S3
Microsoft Azure
Blob Storage
QUERYING
cloudera IMPALA
presto
SQL-ON-HADOOP
ClickHouse
druid
LOW LATENCY
PREPARATION
alteryx
Power BI
tableau
BUSINESS INTELLIGENCE
tableau
Power BI
Looker
BI TOOLS
anodot
outlier
sisu
METRICS MONITORING
Streamlit
Retool
a_
CUSTOM APPS
cube.js
SISENSE
EMBEDDED ANALYTICS
MACHINE LEARNING
alteryx
sas
R
python
DATA SCIENCE PLATFORMS
H2O.ai
DataRobot
AUTO ML PLATFORMS

Data Scientist work at Business Focused/Enterprise

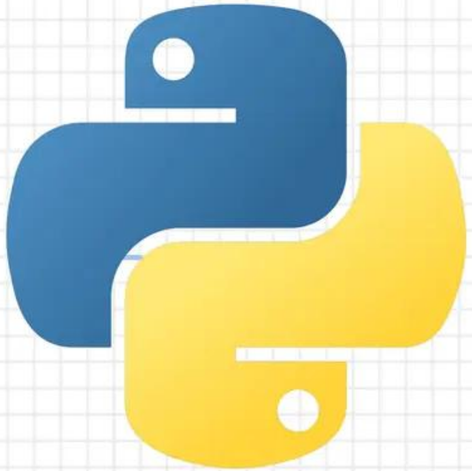


AI Engineering Not Only About The Tech Stack

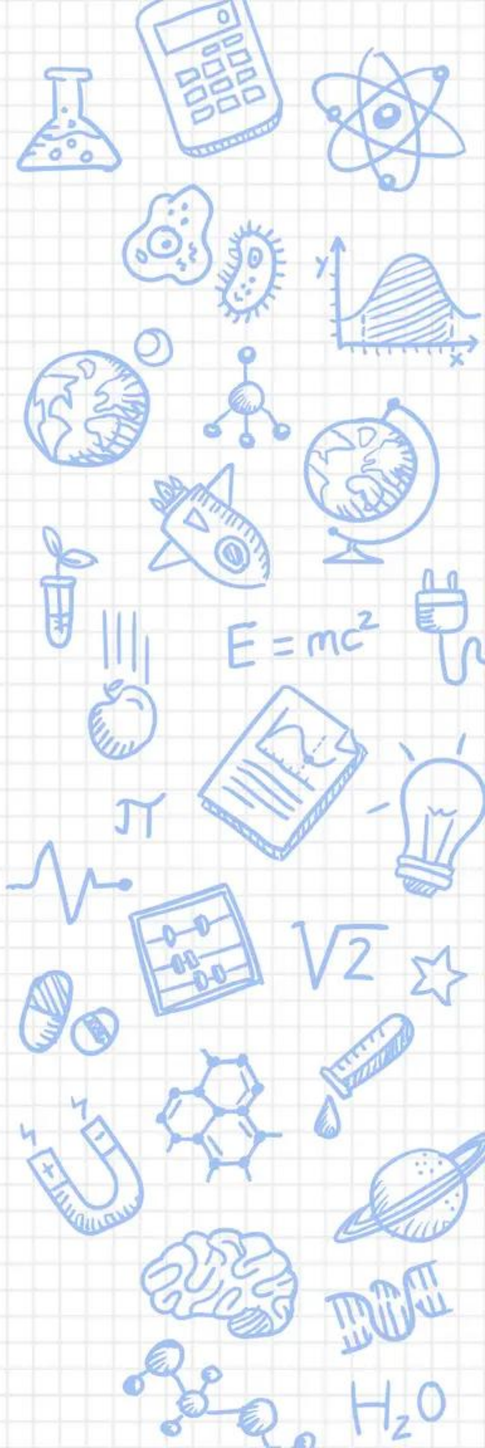
- ✘ SOLID, Design Pattern, Clean Architecture, etcs
- ✘ Software Engineering
- ✘ Tech Tools Characteristics, Behaviour
- ✘ Basic statistics, probilistics and random variable, basic machine learning.
- ✘ High performance computing
- ✘ Security and Regulation
- ✘ ML Ops



The Benefit of Python

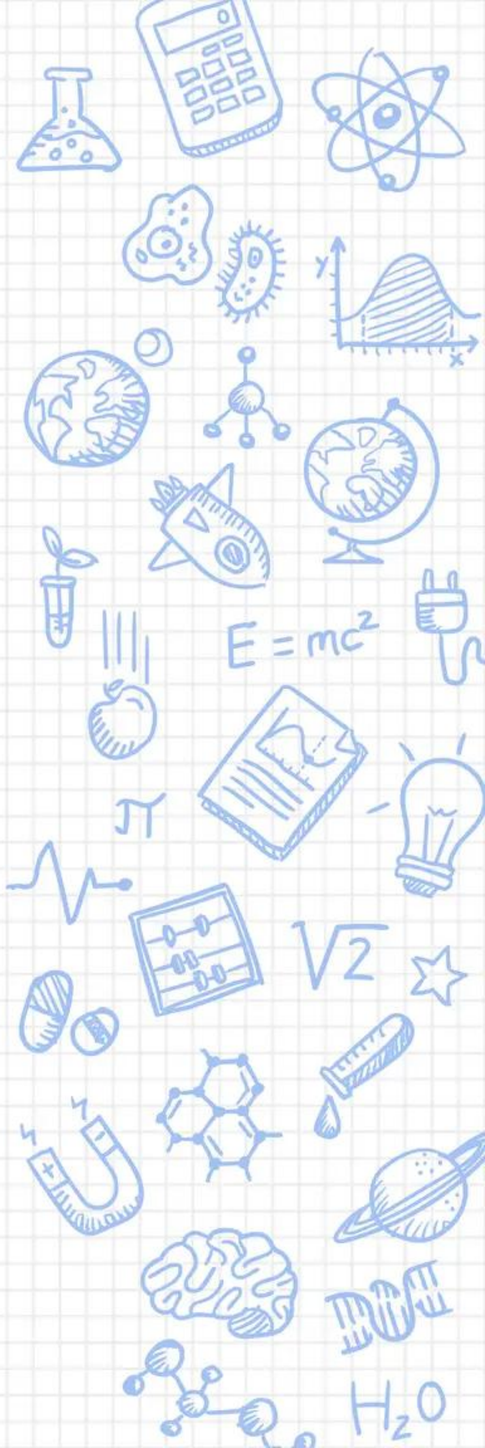


- ✘ Could run AI Algorithm as service
- ✘ Experiments for research phase
- ✘ Having a lot of Math (statistics, probabilistics) and machine learning libraries
- ✘ Data pipeline
- ✘ Easy to use and easy to learn
- ✘ Could implement OOP or FP

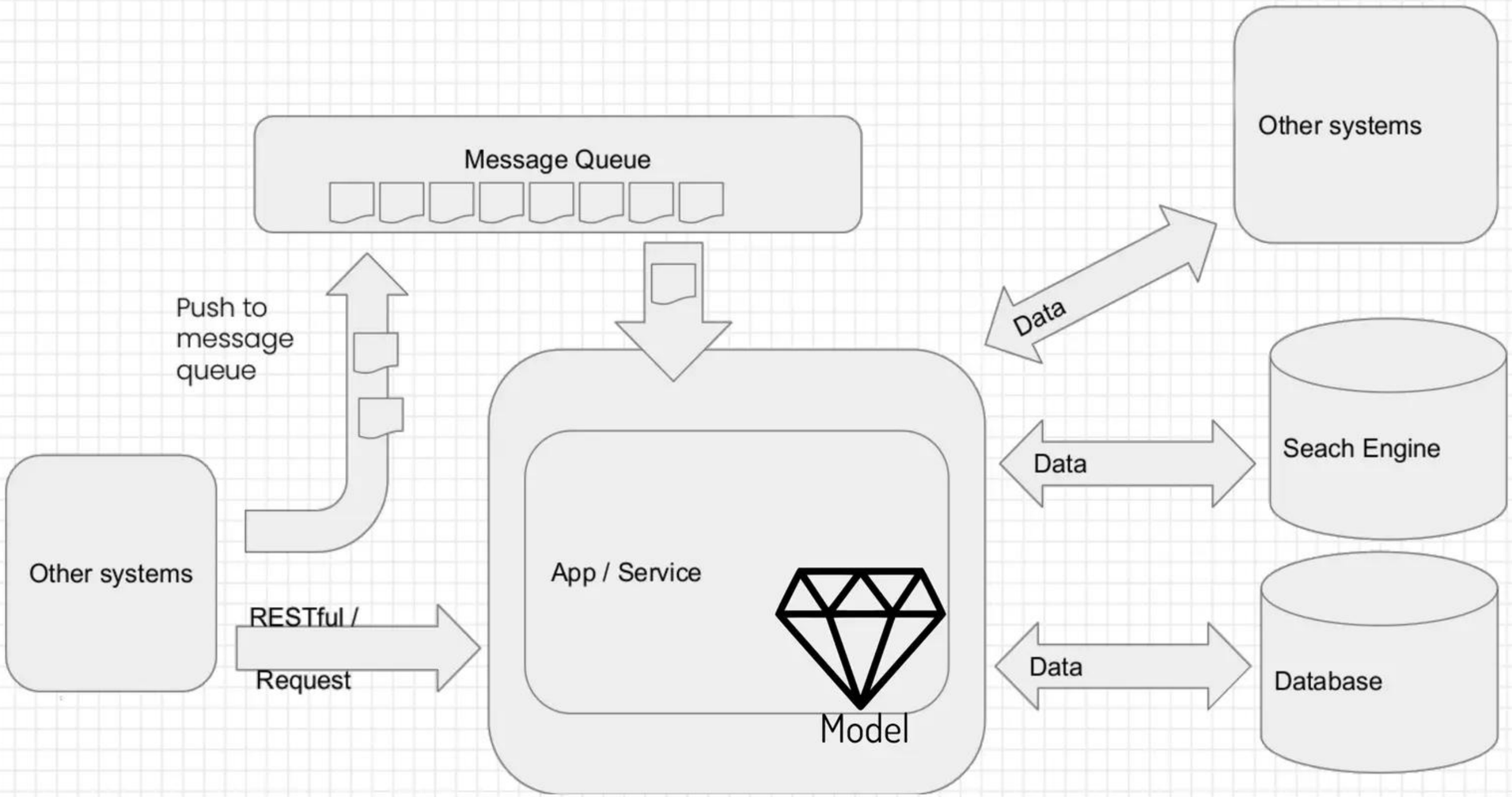


Example of The Engine

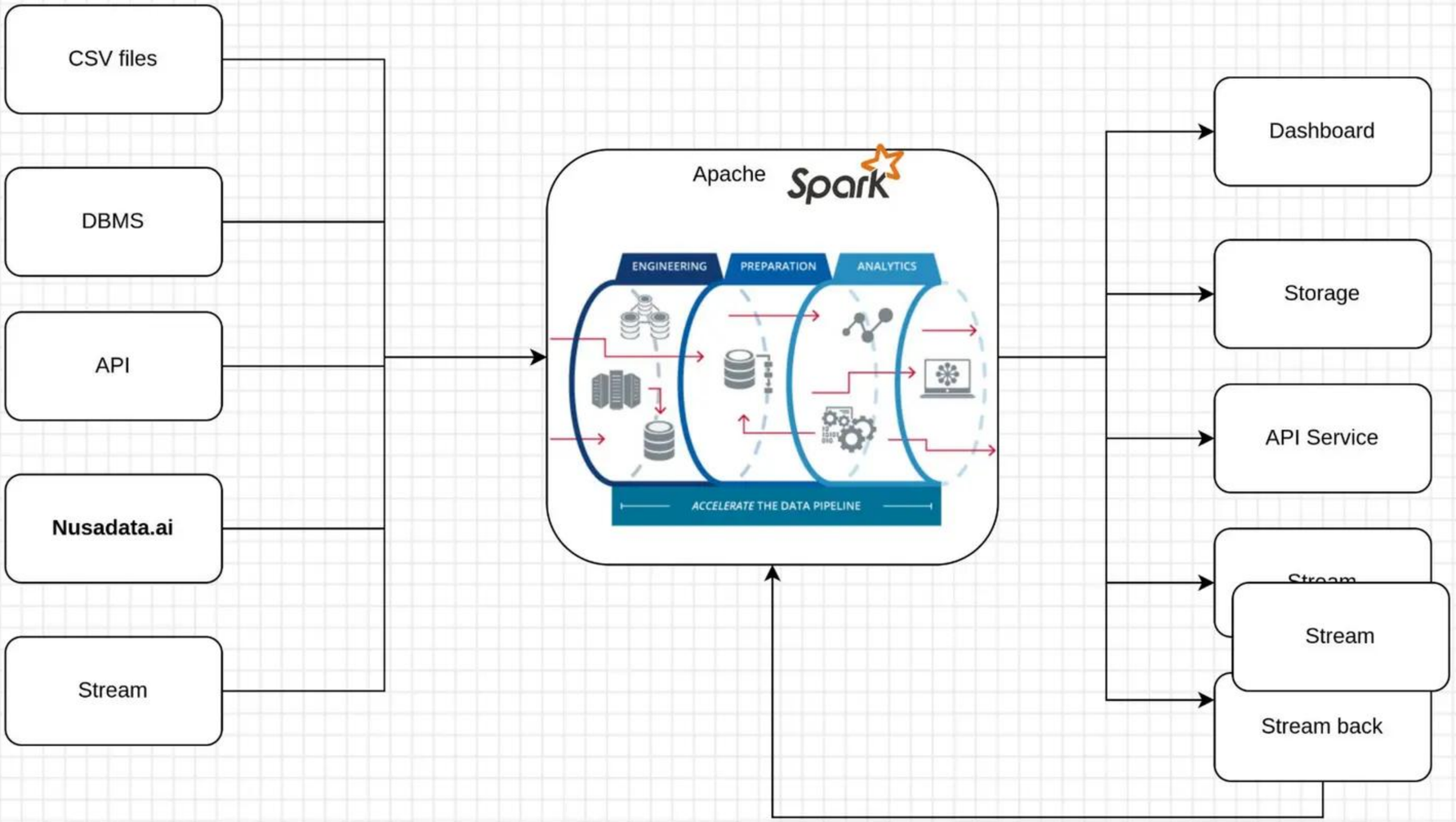
- ✘ We need to run ML service
  - ✗ Sometimes for specific cases
  - ✗ The classification process is about 500ms
- ✘ We run the process (preprocess to main analytics process) as a batch process or need to process data with Large Volume
- ✘ The model is lightweight



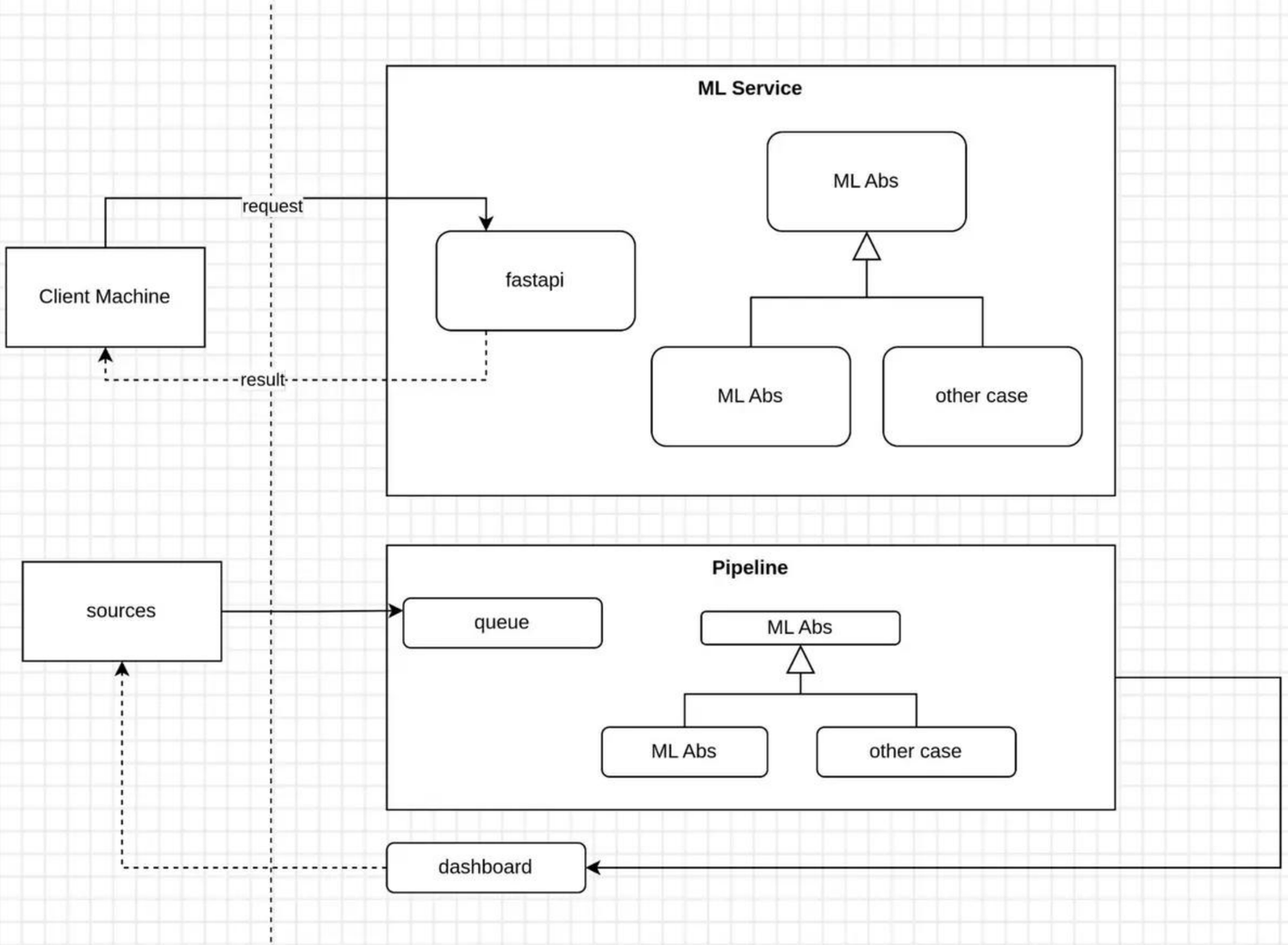
ML Service


Message Queue
Other systems
Push to message queue
Data
Other systems
Seach Engine
Data
App / Service
Model
RESTful / Request
Data
Database

Data batch processing



Could handling several cases

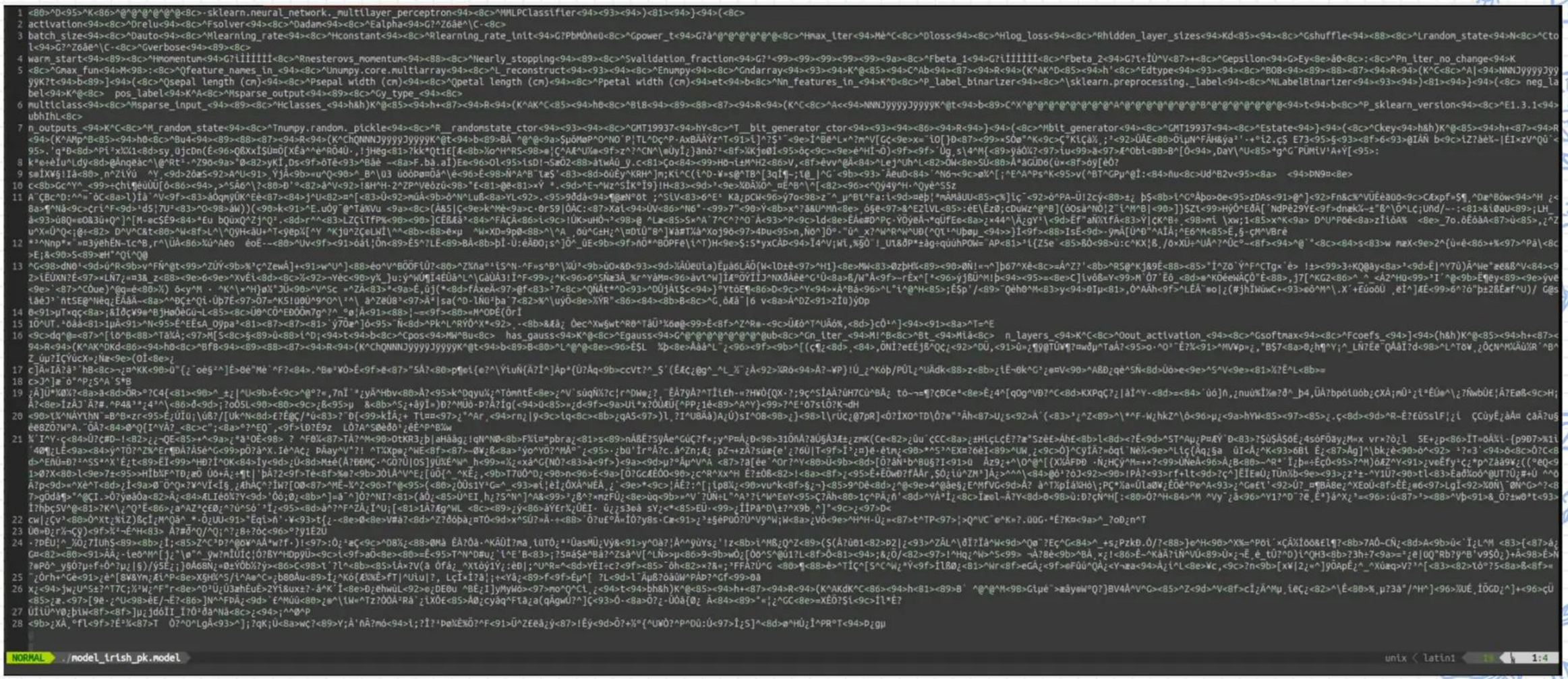


Model

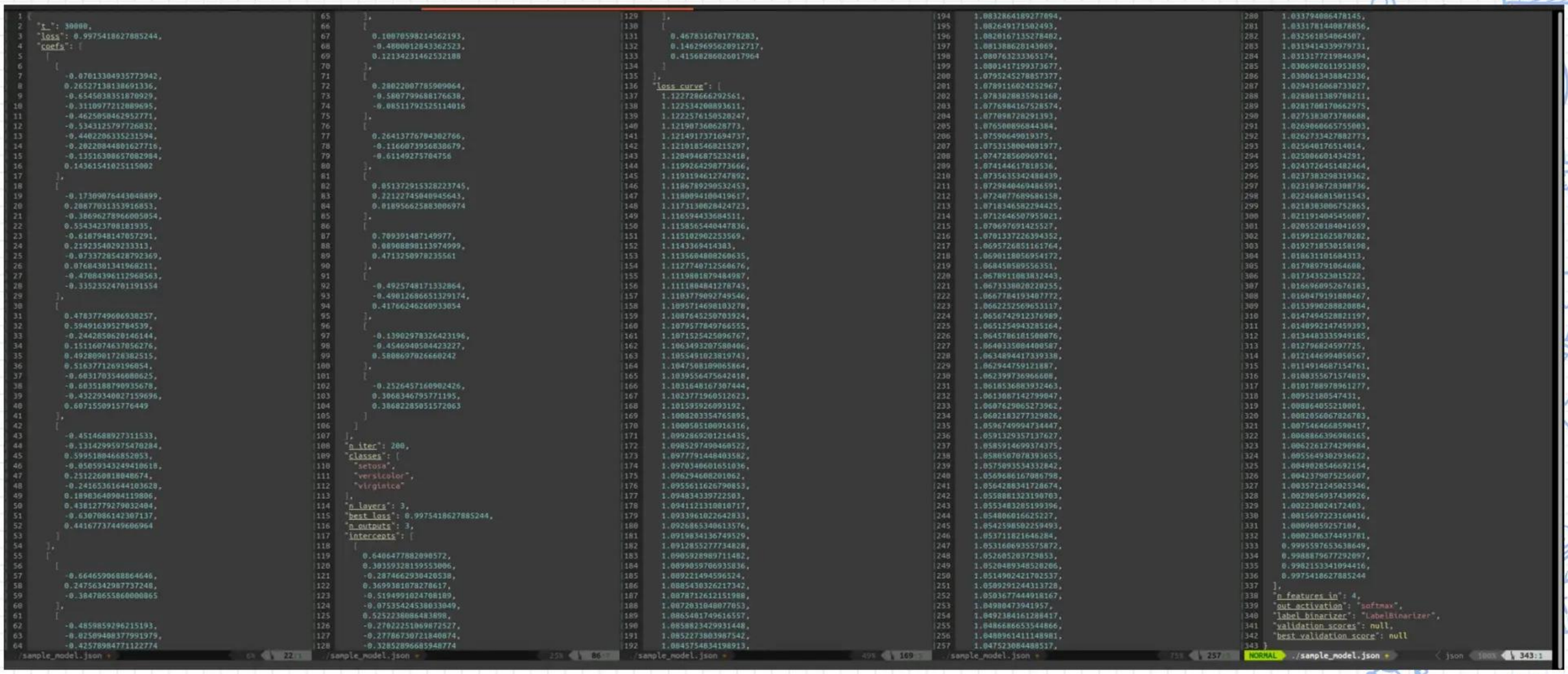
- ✘ We often use pickle or serializable file as model output
- ✘ The size bigger than it should be
- ✘ Impacted to the memory or the space complexity

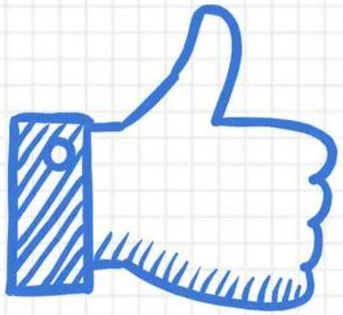
```
total 44K
-rw-rw-r-- 1 hendri hendri 7,7K Okt 22 08:24 model_irish_json.json
-rw-rw-r-- 1 hendri hendri  25K Okt 22 08:41 model_irish_pk.model
-rw-rw-r-- 1 hendri hendri  778 Okt 22 08:41 sample_irish.py
drwxrwxr-x 5 hendri hendri 4,0K Okt 22 08:27 venv
```

Sample Model



Sample Model





THANKS!

Any questions?

You can find me at

- X @infoHendri
- Telegram @siganteng
- Email:

  situkangsayur@gmail.com

  hendri.karisma@akarintidata.ai