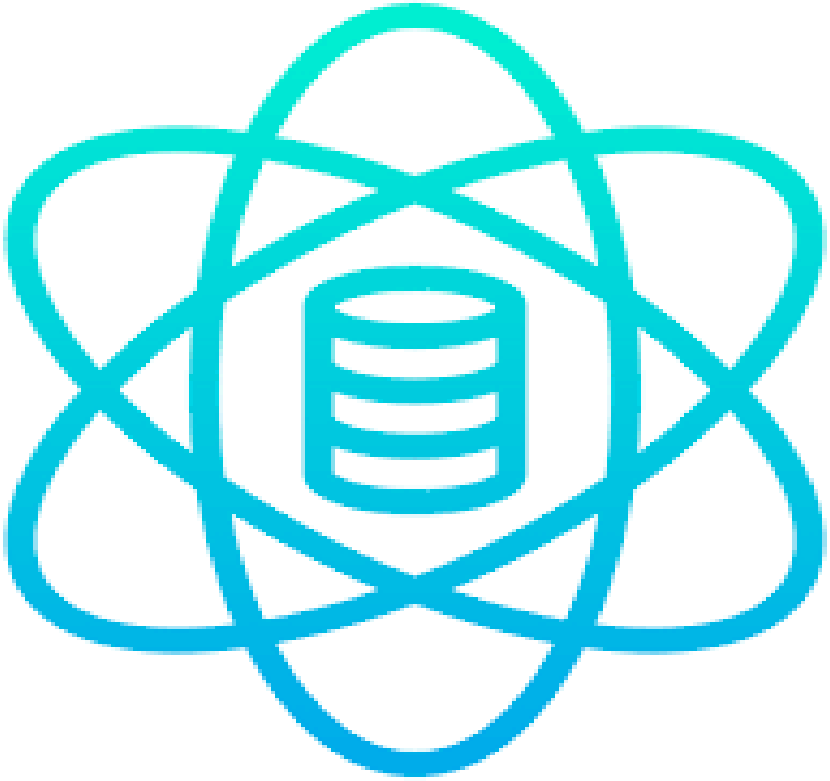
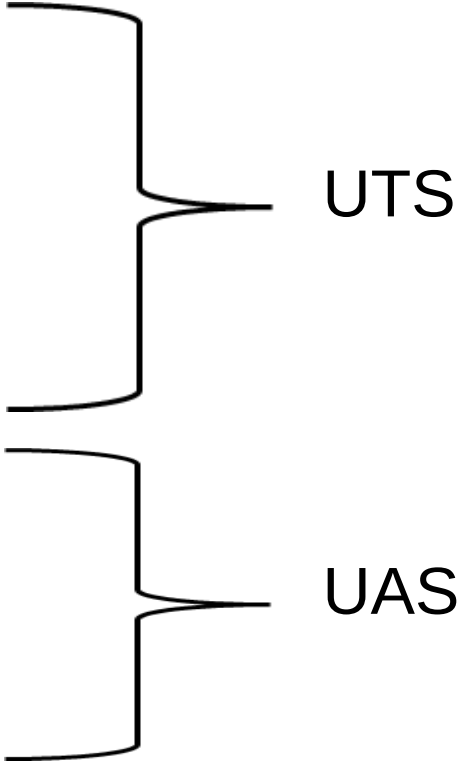


PENGANTAR SAINS DATA

Silabus

1. Pendahuluan Sains Data
2. Data dan Data set
3. Metodologi Sains Data
4. Ekosistem Sains Data
5. Statistika di Sains Data
6. Machine Learning
7. Contoh Standard Task Sains Data
8. Storytelling with the data
9. Sains Data pada Data Tidak Terstruktur
10. KYC dan Mapping Task Sains Data
11. Privacy dan Etika di Data
12. Penerapan Sains Data



Sistem Penilaian

Nilai Akhir = 10% Absensi + 25% Tugas +30% UTS + 35% UAS

Daftar Pustaka

- Grus, J. (2019). *Data science from scratch: first principles with python*. O'Reilly Media.
- Kotu, V., & Deshpande, B. (2018). *Data science: concepts and practice*. Morgan Kaufmann.

PENDAHULUAN SAINS DATA

Tim Dosen Pengantar Sains Data

Outline

- Definisi
- Sejarah dan Evolusi
- Penerapan
- Skill-Set
- Karir
- Mitos-mitos

“The world’s most valuable resource is no longer oil, but data.”

—The Economist, 2017

“Data is the new science. Big Data holds the answers.”

—Pat Gelsinger, CEO at VMware, 2012

Definisi Sains Data

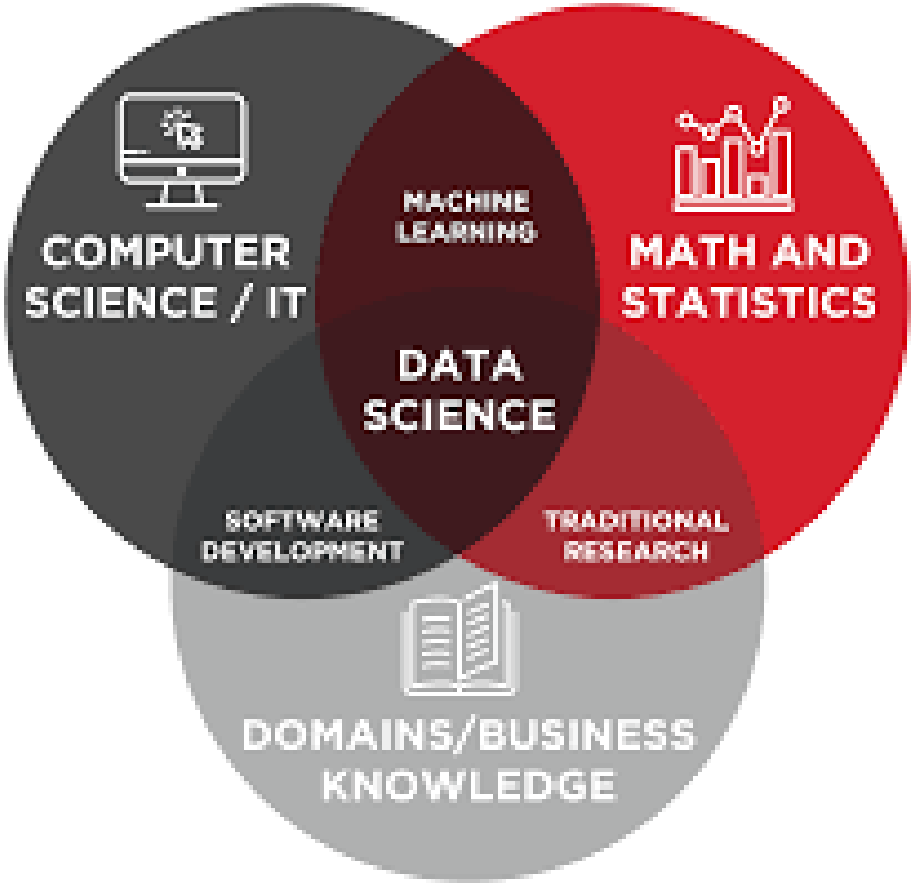
Definisi

Data science adalah kombinasi matematika, statistik, programming, analitik lanjutan, kecerdasan buatan dan pembelajaran mesin dalam bidang keahlian tertentu untuk menemukan *actionable insights* yang tersembunyi dalam kumpulan data suatu organisasi. Insights tersebut dapat digunakan untuk panduan pembuatan keputusan dan perencanaan strategis..

https://www.ibm.com/cloud/learn/data-science-introduction#:

Data Science merupakan keterampilan yang membutuhkan ilmu komputer, pemrograman, teknologi, dan statistik Keterampilan ini mencakup teknologi dan teknik seperti memanfaatkan komputasi Cloud, analisis Big Data, pemrosesan Natural Language, pembelajaran tanpa pengawasan (Unsupervised Learning) seperti analisis Cluster, Web Scraping, teknik Fuzzy, Machine Learning, dan lain sebagainya

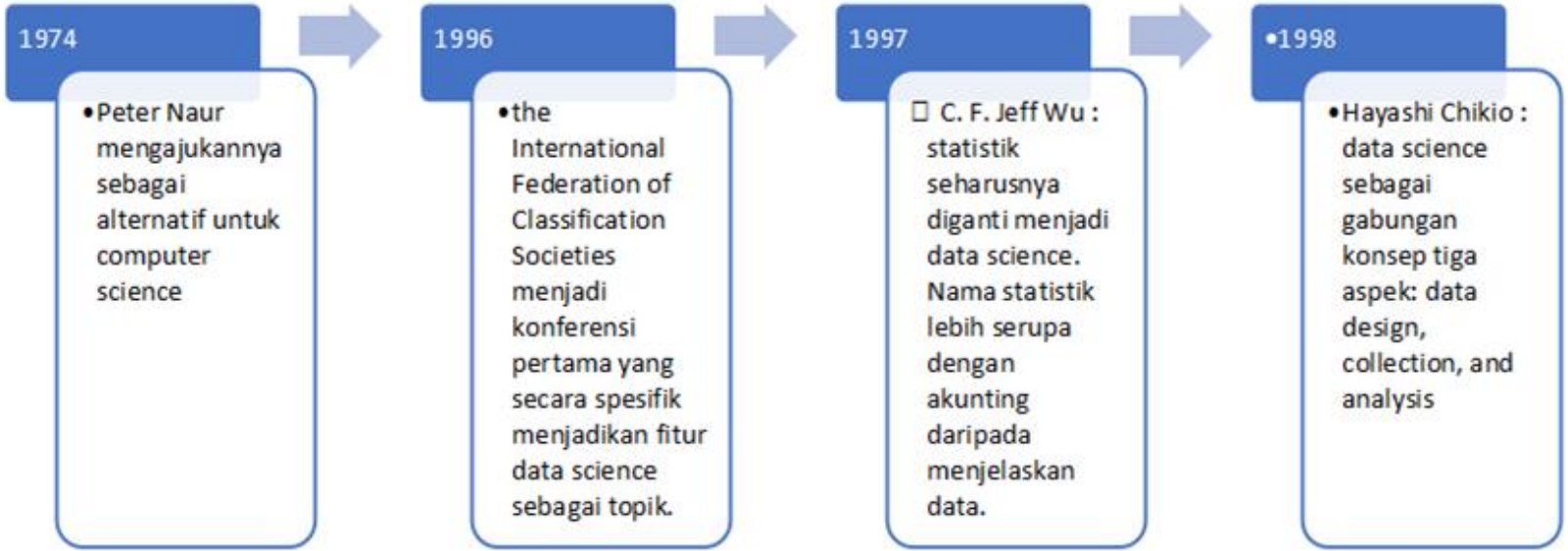
https://www.urban.org/research-methods/data-science


COMPUTER SCIENCE / IT
MACHINE LEARNING
MATH AND STATISTICS
DATA SCIENCE
SOFTWARE DEVELOPMENT
TRADITIONAL RESEARCH
DOMAINS/BUSINESS KNOWLEDGE

Sejarah & Evolusi Sains Data

Penggunaan Awal Istilah Data Science

Penggunaan istilah Data Science Modern



A Brief History of Data Science

1962

John Wilder Tukey
expounded on a move in the world of statistics in 1962.

1974

Peter Naur
wrote the Concise Survey of Computer Methods, utilizing the expression "Data Science," over and over.

1977

John Wilder Tukey
composed a second paper titled "Exploratory Data Analysis"

1989

1994

Gregory Piatetsky — Shapiro
arranged a Knowledge Discovery in Databases workshop.

Business Week ran the main story.
Database Marketing

1999

2001

Jacob Zahavi brought up the requirement for new devices to deal with the enormous measures of data accessible to organizations.

International Council for Science:
Committee on Data for Science and Technology started distributing the **Data Science Journal**

2006

2008

Hadoop 0.1.0,
an open-source and non-relational database was released

2011

Title "Data Scientist"
turned into a trendy expression and in the long run a piece of the language.

Data Scientists
expanded by 15,000% in between 2011 – 2012.

DatabaseTown.com

Penerapan Sains Data

GOVERNMENT

Some of applications in energy exploration, financial market analysis, etc.

HEALTHCARE

Allows faster identification and efficient analysis of healthcare information.

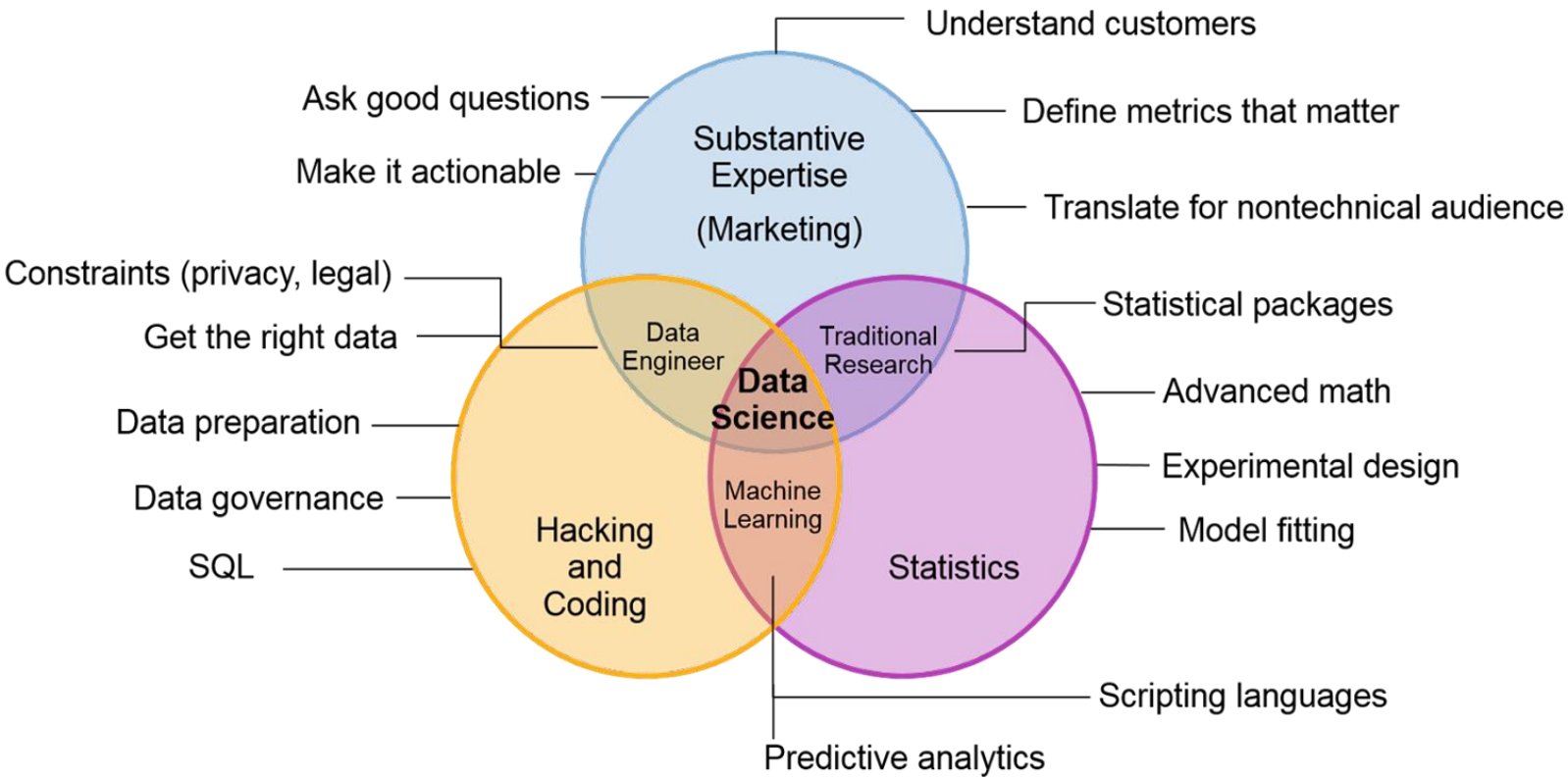
BANKING

Card fraud detection, enterprise credit risk reporting, etc.

RETAIL

Predicting spending, personalizing customer experience, etc.

Skill Sains Data


Understand customers
Ask good questions
Define metrics that matter
Substantive Expertise (Marketing)
Make it actionable
Translate for nontechnical audience
Constraints (privacy, legal)
Statistical packages
Get the right data
Data Engineer
Traditional Research
Data Science
Advanced math
Data preparation
Experimental design
Data governance
Machine Learning
Model fitting
Hacking and Coding
Statistics
SQL
Scripting languages
Predictive analytics

Data Science Learning Path

Data Scientist

Roadmap

Mathematics

- Linear Algebra
- Analytics Geometry
- Matrix
- Vector Calculus
- Optimization
- Regression
- Dimensionality Reduction
- Density Estimation
- Classification

Probability

- Discrete Distribution
  - *Binomial*
  - *Bernoulli*
  - *Geometric etc*
- Continuos Distribution
  - *Uniform*
  - *Exponential*
  - *Gamma*
- Normal Distribution
- Introduction to Probability
- 1D Random Variable
- Function of One Random Variable
- Joint Probability Distribution

Statistics

- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypotheses Testing
- ANOVA
- Reliability Engineering
- Stochastic Process
- Computer Simulation
- Design of Experiments
- Simple Linear Regression
- Correlation
- Multiple Regression
- Nonparametric Statistics
  - *Sign Test*
  - *The Wilcoxon Signed-Rank Test*
  - *The Wilcoxon Rank Sum Test*
  - *The Kruskal-Walls Test*
- Statistical Quality Control
- Basic of Graphs

Programming

Python

- Python Basics
  - *List*
  - *Set*
  - *Tuples*
  - *Dictionary*
  - *Function, etc.*
- NumPy
- Pandas
- Matplotlib/Seaborn, etc.

R

- R Basic
  - *Vector*
  - *List*
  - *Data Frame*
  - *Matrix*
  - *Array, etc*
- dplyr
- ggplot2
- Tidyr
- Shiny, etc.

DataBase

- SQL
- MongoDB

Other

- Data Structure
  - *Array, etc*
- Web Scraping
- Linux
- Git

Machine Learning

Introduction

- How Model Works
- Basic Data Exploration
- First ML Model
- Model Validation
- Underfitting & Overfitting
- Random Forests
- scikit-learn

Intermediate

- Handling Missing Values
- Handling Categorical Variables
- Pipelines
- Cross-Validation
- XGBoost
- Data Leakage

Deep Learning

- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- PyTorch
- TensorFlow
- A Single Neuron
- Deep Neural Network
- Stochastic Gradient Descent
- Overfitting and Underfitting
- Dropout Batch Normalization
- Binary Classification

Feature Engineering

- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural language Processing

- Text Classification
- Word Vectors

Data Visualization Tools

- Excel VBA
- BI (Business Intelligence)
  - *Tableau*
  - *Power BI*
  - *Qlik View*
  - *Qlik Sense*

Deployment

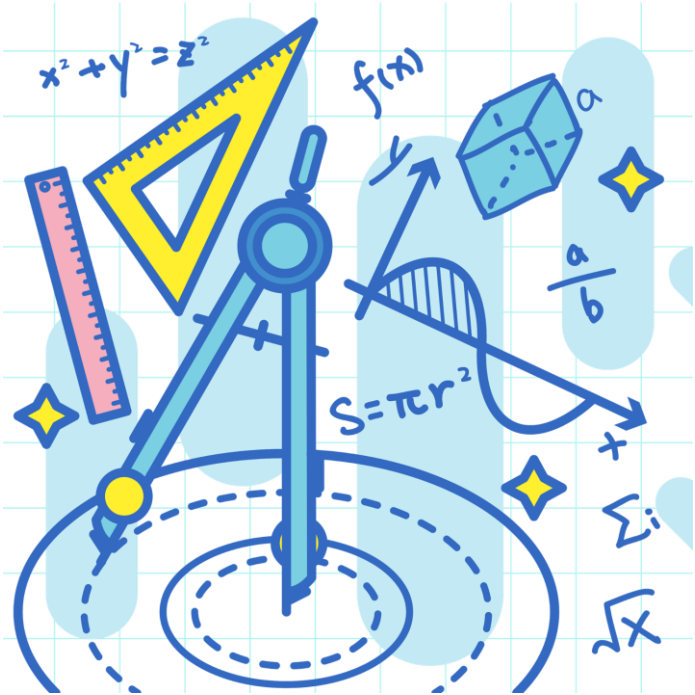
- Microsoft Azure
- Heroku
- Google Cloud Platform
- Flask
- Django

Other Points

- Domain Knowledge
- Communication Skill
- Reinforcement Learning
- Case Studies
  - *Data Science at Netflix*
  - *Data Science at Flipkart*
  - *Project on Credit Card Fraud Detection*
  - *Project on Movie Recommendation , etc.*

Keep Practicing

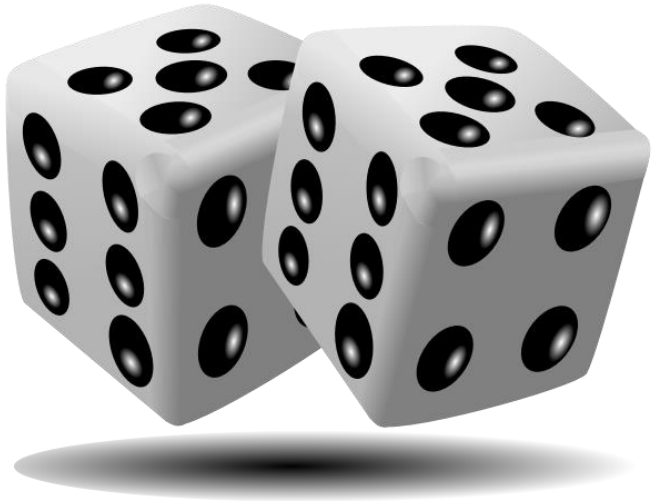
Mathematics



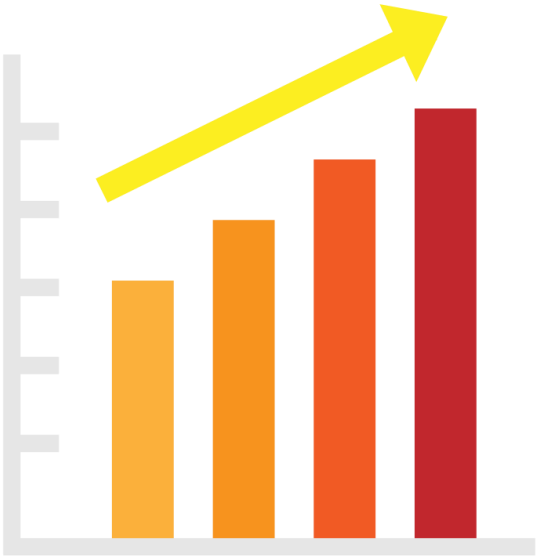
- Linear Algebra
- Analytics Geometry
- Matrix
- Vector Calculus
- Optimization
- Regression
- Dimensionality Reduction
- Density Estimation
- Classification

Probability

- Discrete Distribution (Binomial, Bernouli, Geometric, dll)
- Continuous Distribution (Uniform, Exponential, Gamma)
- Normal Distribution
- Introduction to Probability
- 1D Random Variable
- Function of One Random Variable
- Joint Probability Distribution



Statistics

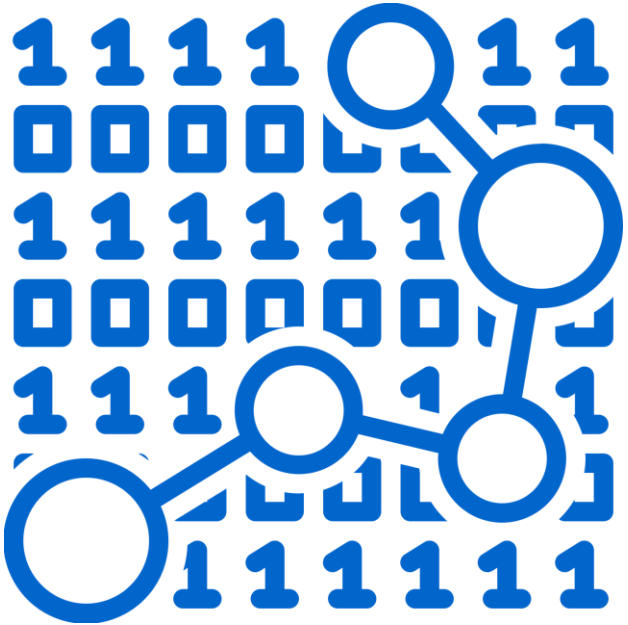


- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypothesis Testing
- Correlation
- Multiple Regression
- Basic of Graphs
- Computer Simulation

Programming

- Python (Basics, NumPy, Pandas, Jupyter Notebook, etc)
- R (Basics, dplyr, ggplot2, Tidyr, Shiny, etc)
- Database (MongoDB, SQL)
- Data Structures
- Web Scraping
- Git
- etc

Machine Learning

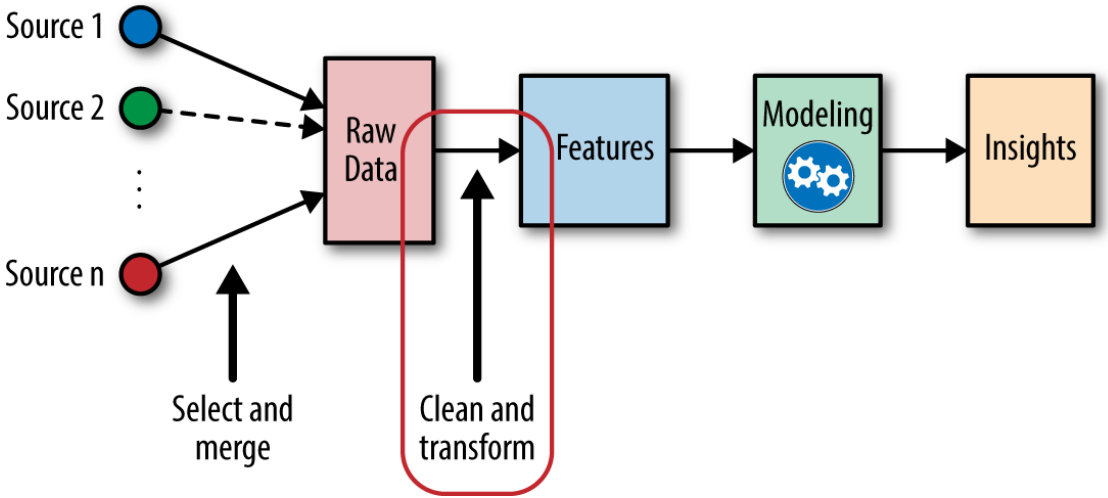


- How Model Works
- Basic Data Exploration
- ML Model
- Model Validation
- Underfitting/Overfitting
- Random Forests
- Pipelines
- Cross-Validation
- Data Leakage
- etc

Deep Learning

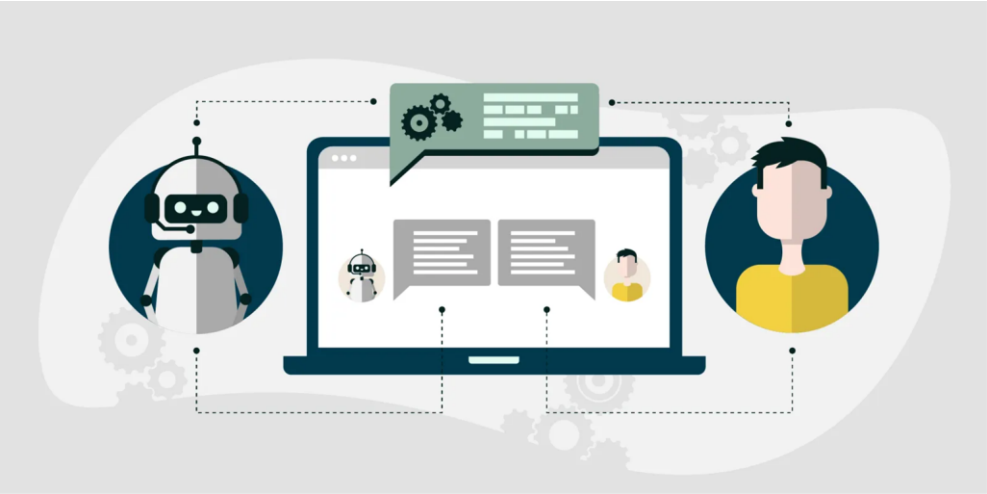
- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- Keras
- PyTorch
- TensorFlow
- A Single Neuron
- Deep Neural Network
- Binary Classification
- etc

Feature Engineering



- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural Language Processing



- Text Classification
- Word Vectors

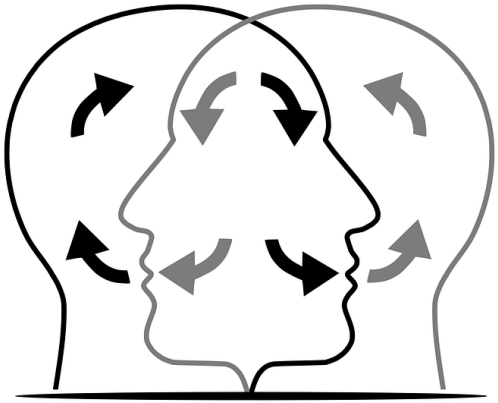
Data Visualization Tools

- Tableau
- Power BI
- Google Data Studio
- Data Wrapper
- Visme
- etc

Deployment

- Microsoft Azure
- Google Cloud Platform
- Django
- Heroku

Other Points (Soft Skills)



- Domain Knowledge
- Communication Skill

Karir di Sains Data

Data Analyst

Data Quality Engineer

Database Administrator

Data Modeler

BI Engineer

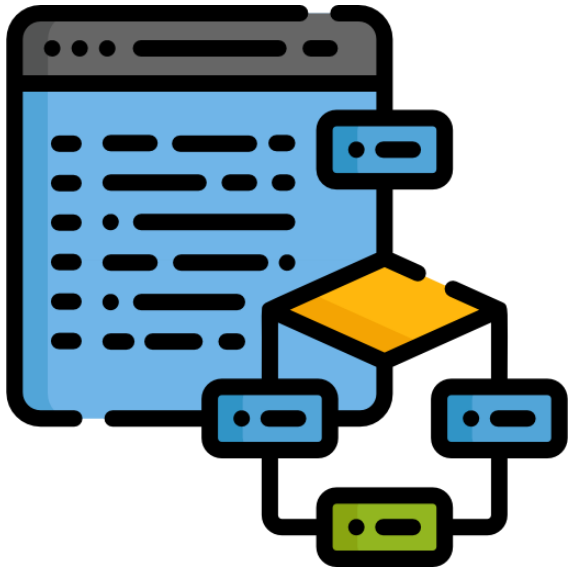
Data Engineer

Data Scientist

Data Architect

Mitos - mitos Seputar Sains Data

Ahli Sains Data harus ada di semua perusahaan



Ahli Sains Data lebih banyak mengurusi algoritma dan hal - hal rumit lainnya

Ahli Sains Data harus bergelar doktor (Ph.D)

Semua orang bisa meniti karier di Sains Data


Terima
Kasih