

# Metodologi Data Science dan Bisnis Data Science

Pertemuan ke-1

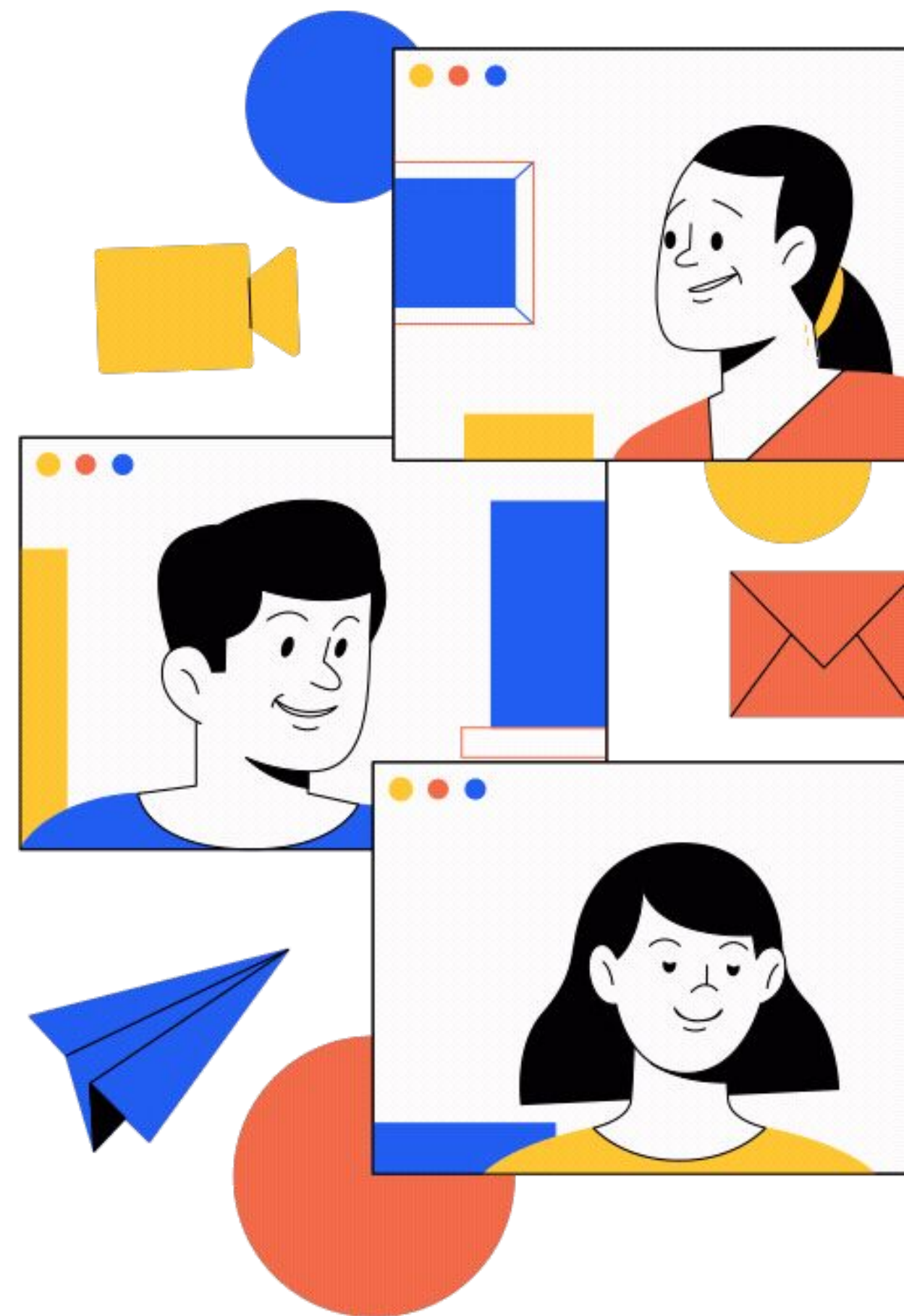
## Nama Pengajar

Dian Ade Kurnia, M.Kom

Rusnanda Farhan

Rusnandi Fikri

Rika Sahriana



# Tujuan Pembelajaran

01

Membahas metodologi data science secara umum

02

Mengembangkan aplikasi AI dengan menjelaskan Langkah-Langkah utama yang diperlukan untuk menyelesaikan masalah organisasi/bisnis dengan melakukan tugas-tugas yang terkait dengan data science



# Capaian Pembelajaran

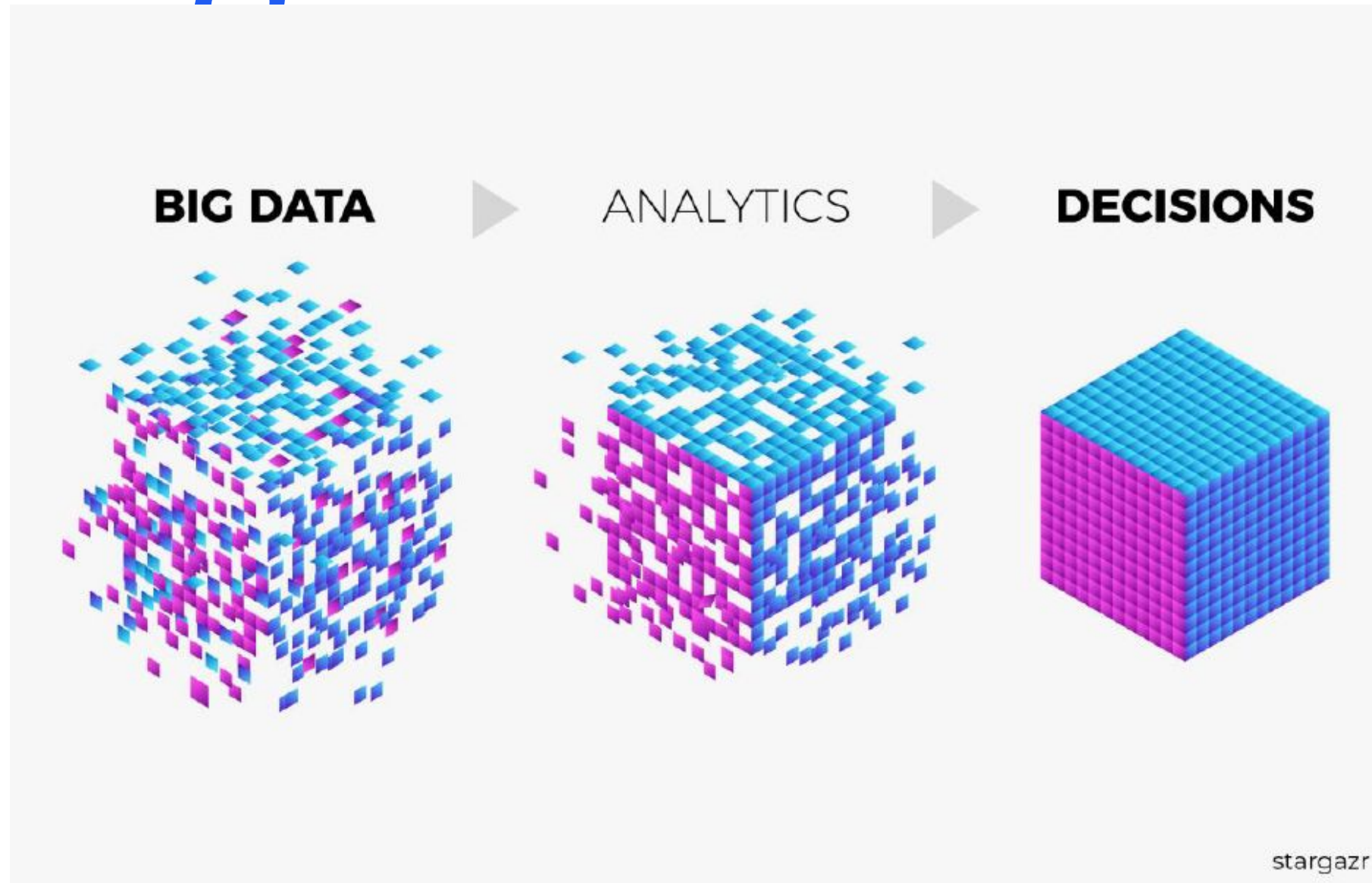
**Yang akan kita bahas pada sesi ini**

- Metodologi Data Science
- Langkah-Langkah Utama dalam Metodologi Data Science





# Sistem AI berbasis Big Data



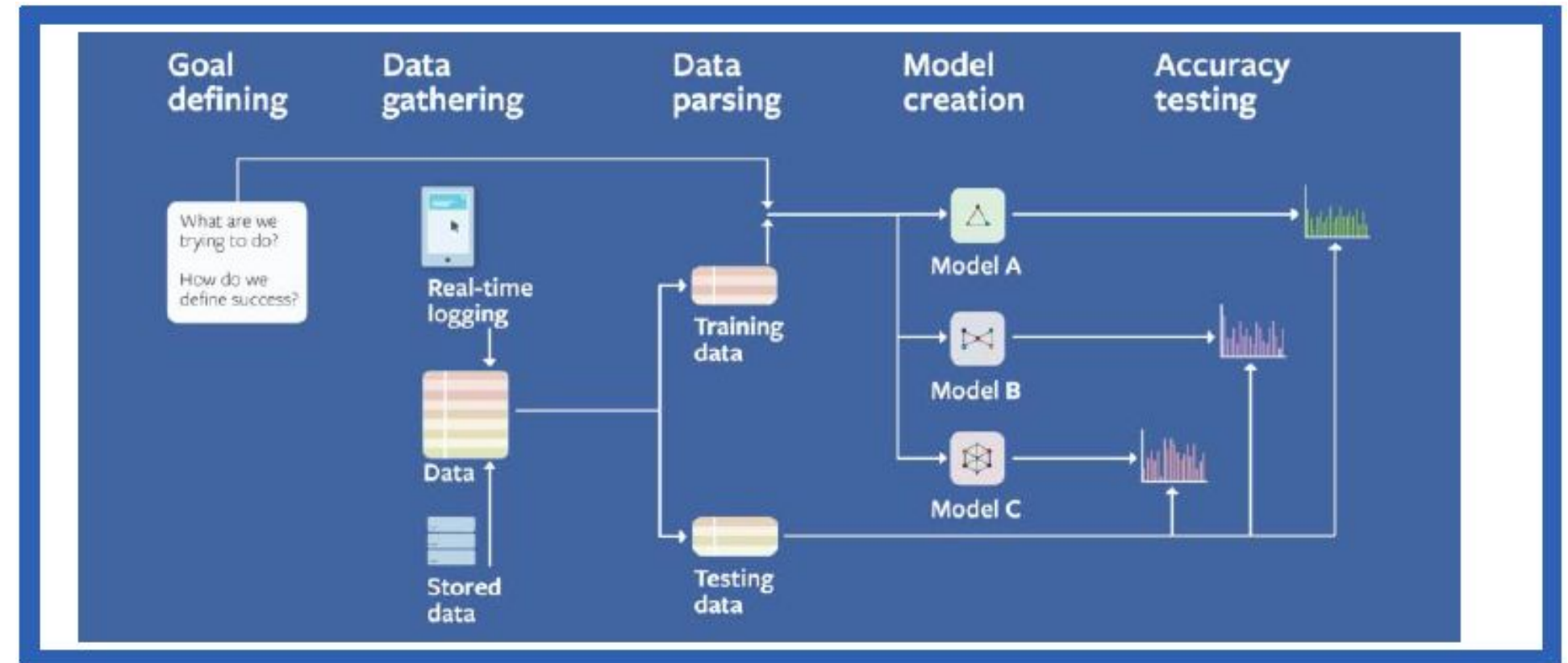
**DATA**



**INTELLIGENCE SYSTEM  
(KNOWLEDGE BASE)**

# Tahapan Pengembangan Sistem AI berbasis Big Data

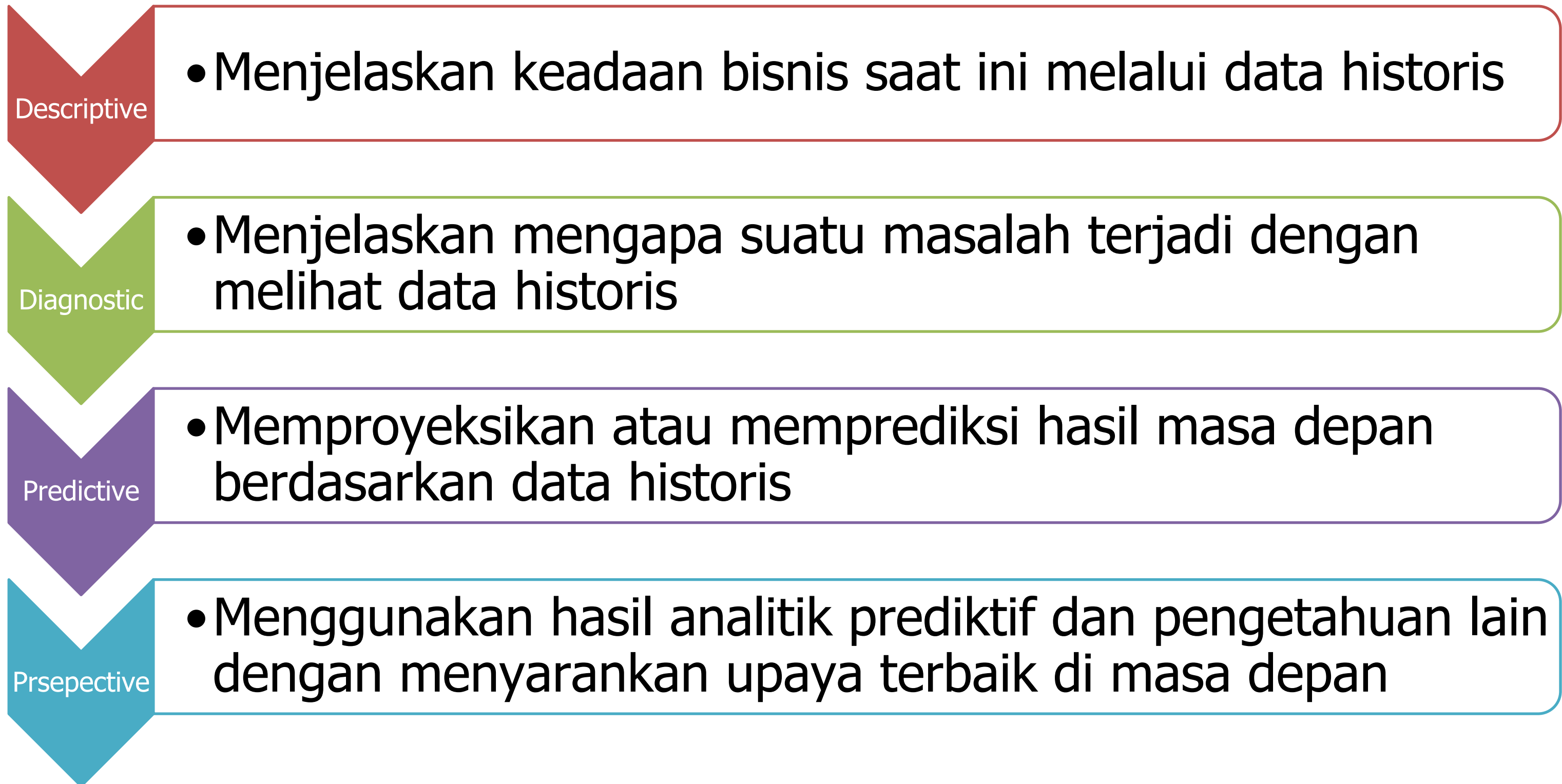
Pengembangan melalui Pelatihan (TRAINING)



Penggunaan



# Tugas (Task) apa yang biasa dikembangkan



# Jenis Task yang dikembangkan

Regresi

Classification

Clustering

Association

Anomaly  
Detection

Sequence Mining

Recommendation  
System





# Metodologi Data Science



# Jenis Metodologi



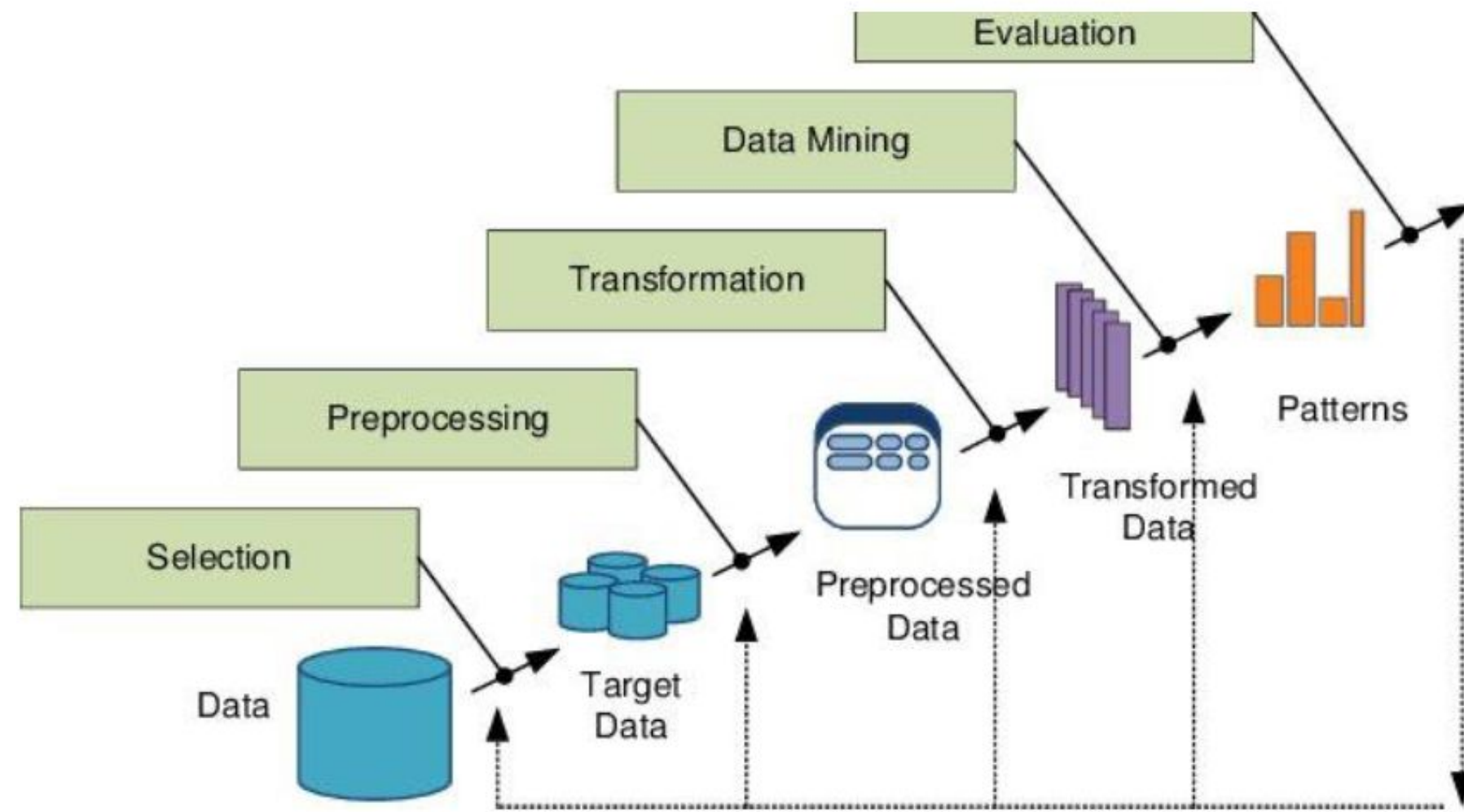
```
graph TD; A[Jenis Metodologi] --- B(( )); A --- C(( )); B --- D[Metodologi Kegiatan Teknis]; C --- E[Metodologi Kegiatan Bisnis (dan Teknis)];
```

Metodologi Kegiatan Teknis

Metodologi Kegiatan Bisnis (dan Teknis)

# Metodologi Teknis : Kegiatan DS / AI dianggap Kegiatan Teknikal

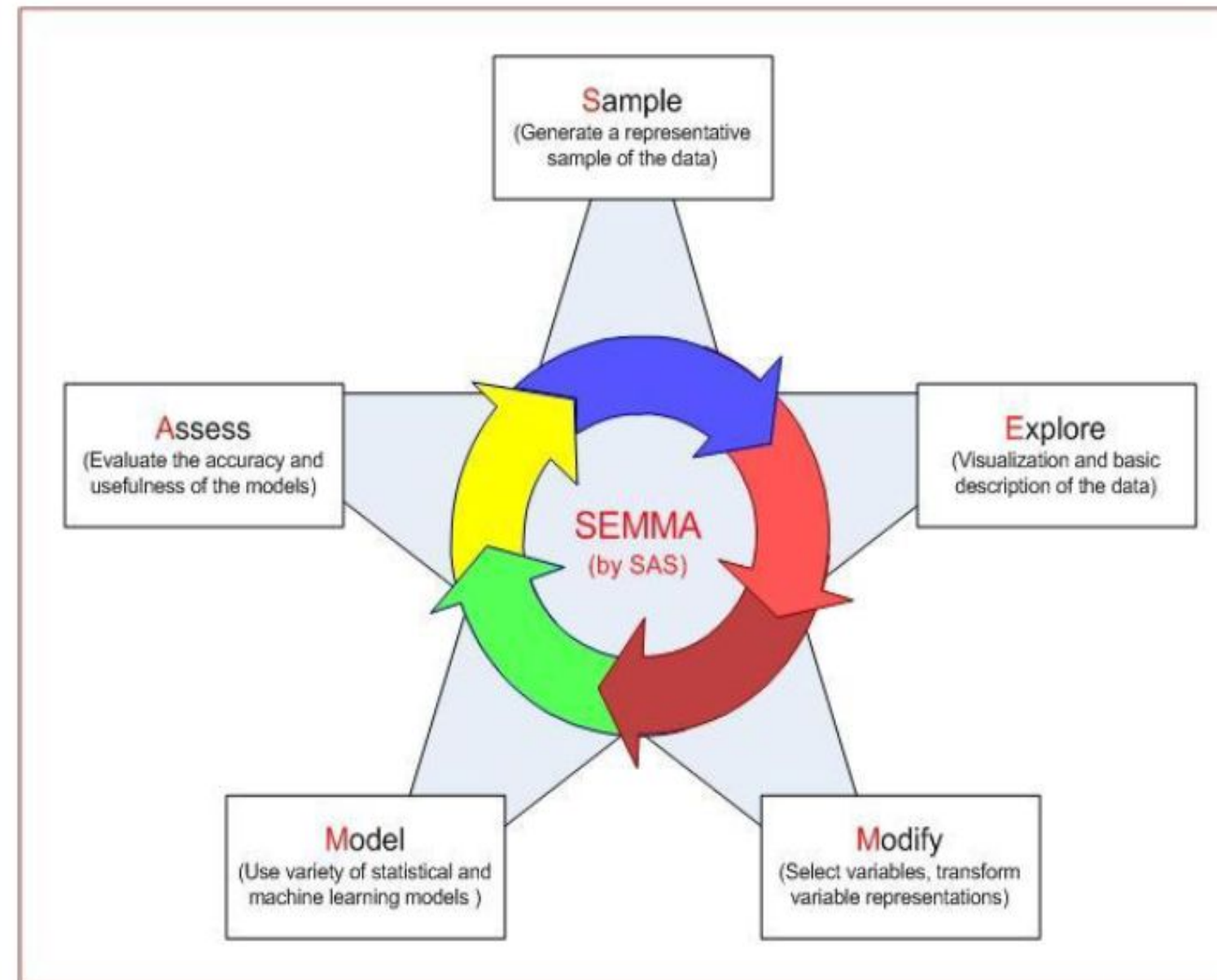
Knowledge Discovery and Data Mining  
(KDD)



<https://www.kdnuggets.com/gpspubs/ai-mag-kdd-overview-1996-Fayyad.pdf>

# Metodologi Teknis : Kegiata DS / AI dianggap Kegiatan Teknikal

SEMMA  
Dari SAS Institute

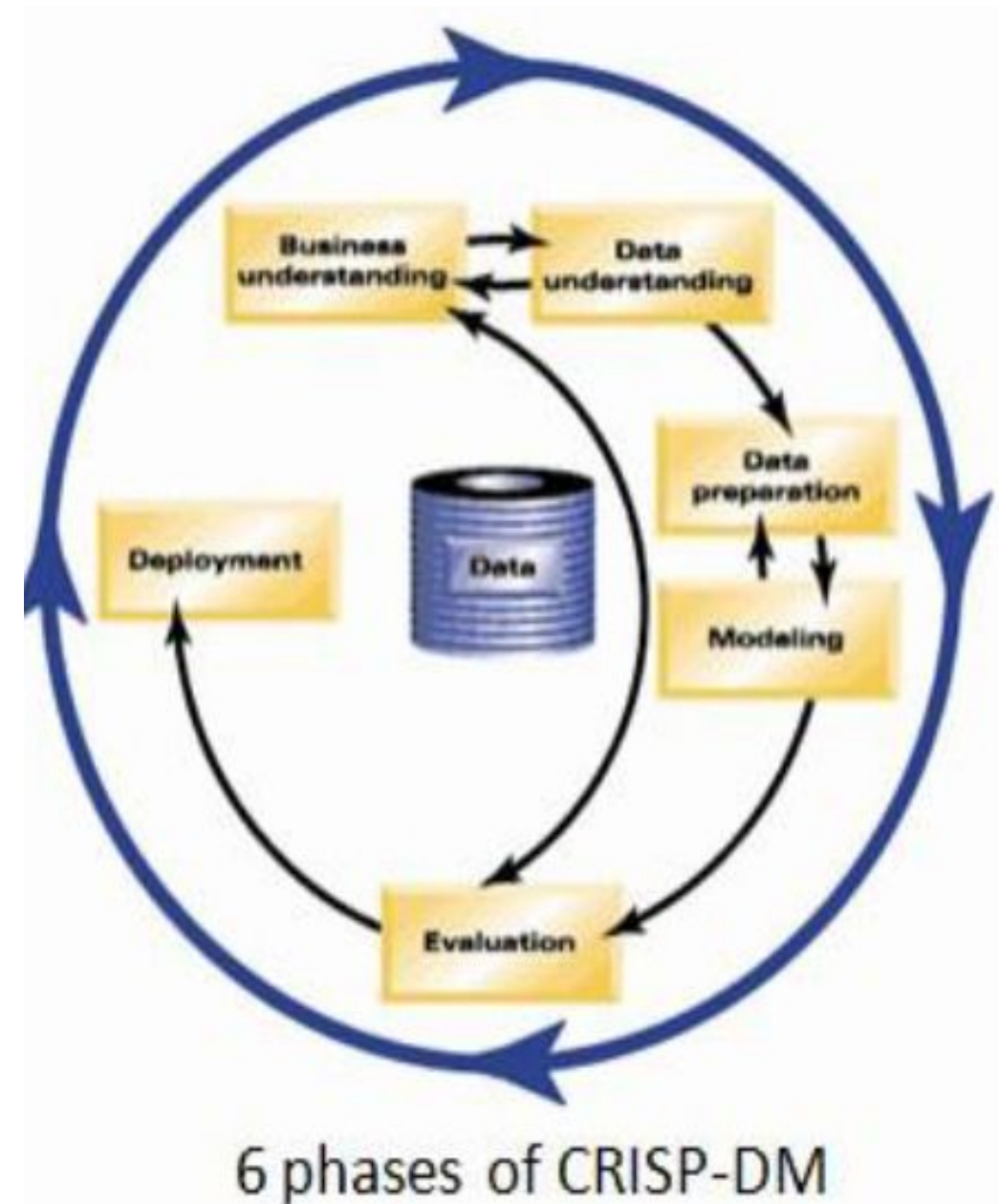


<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbj1a2.htm&docsetVersion=14.3&locale=en>



# Metodologi Lengkap : Kegiata DS / AI dianggap Kegiatan Bisnis : Masalah Bisnis menjadi masalah DS/AI

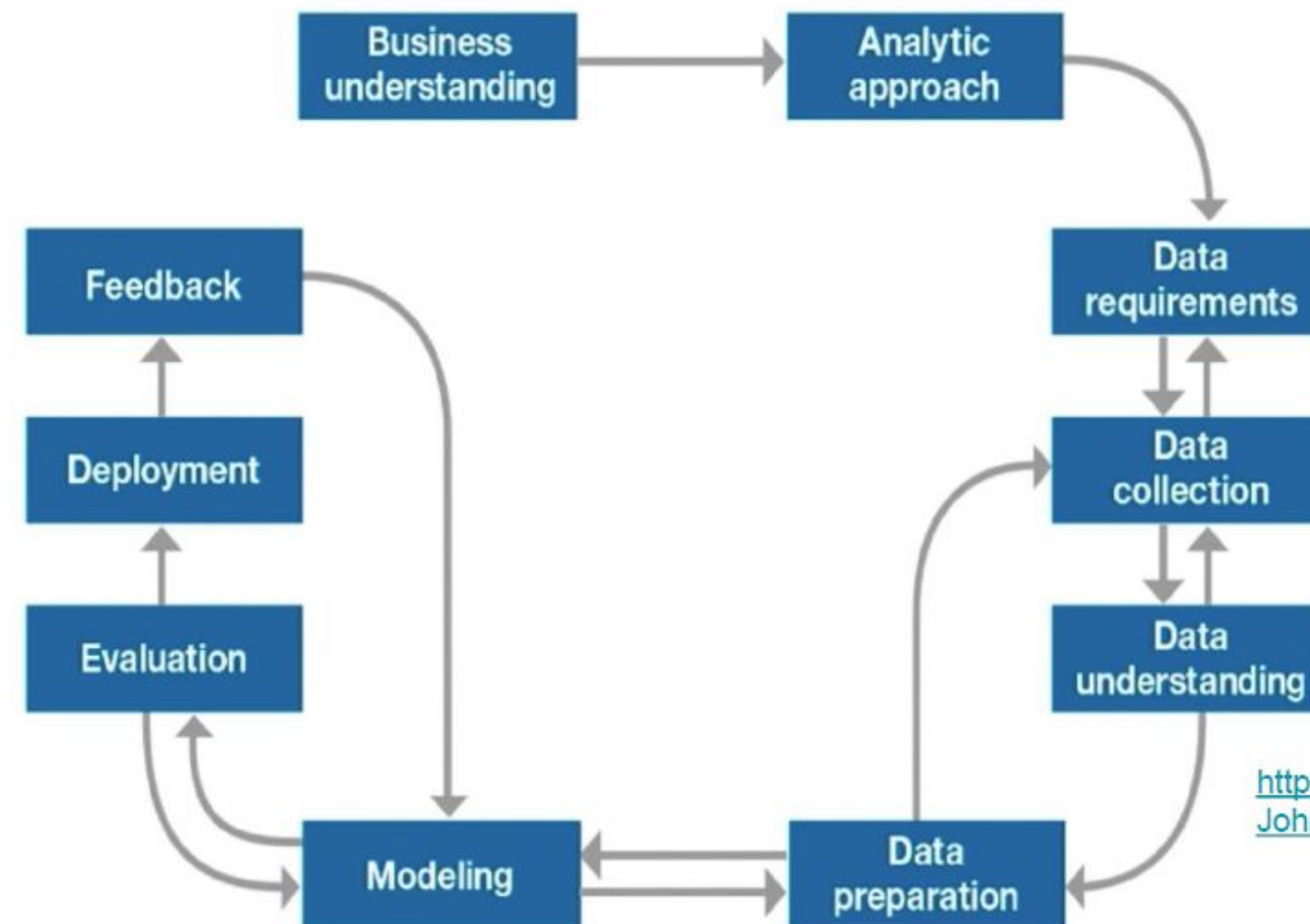
Cross-Industry Standard Process for Data Mining (CRISP-DM)



<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en>

# Metodologi Lengkap : Kegiatan DS / AI dianggap Kegiatan Bisnis : Masalah Bisnis menjadi masalah DS/AI

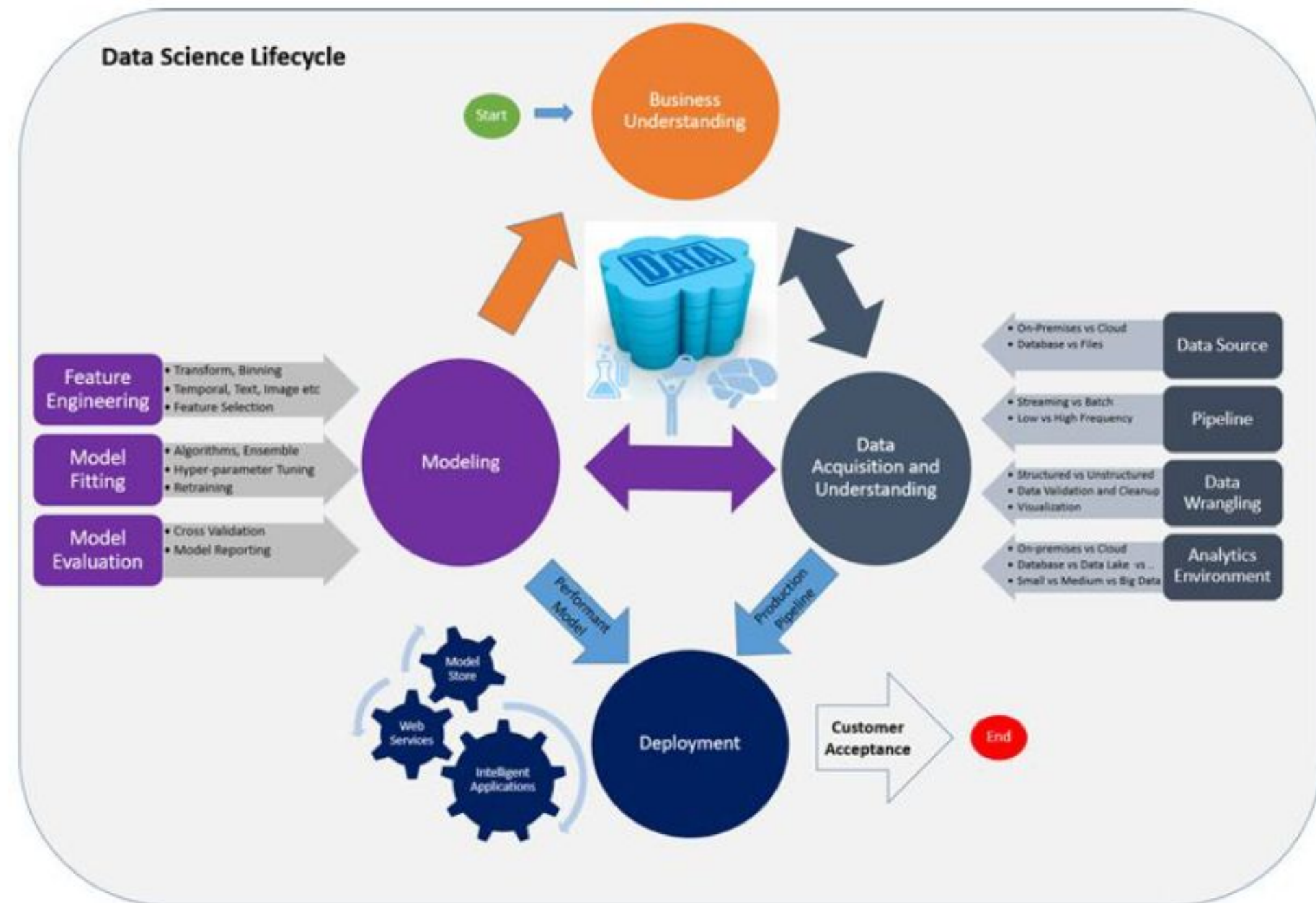
IBM Data Science  
Methodology



<https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science>

# Metodologi Lengkap : Kegiatan DS / AI dianggap Kegiatan Bisnis : Masalah Bisnis menjadi masalah DS/AI

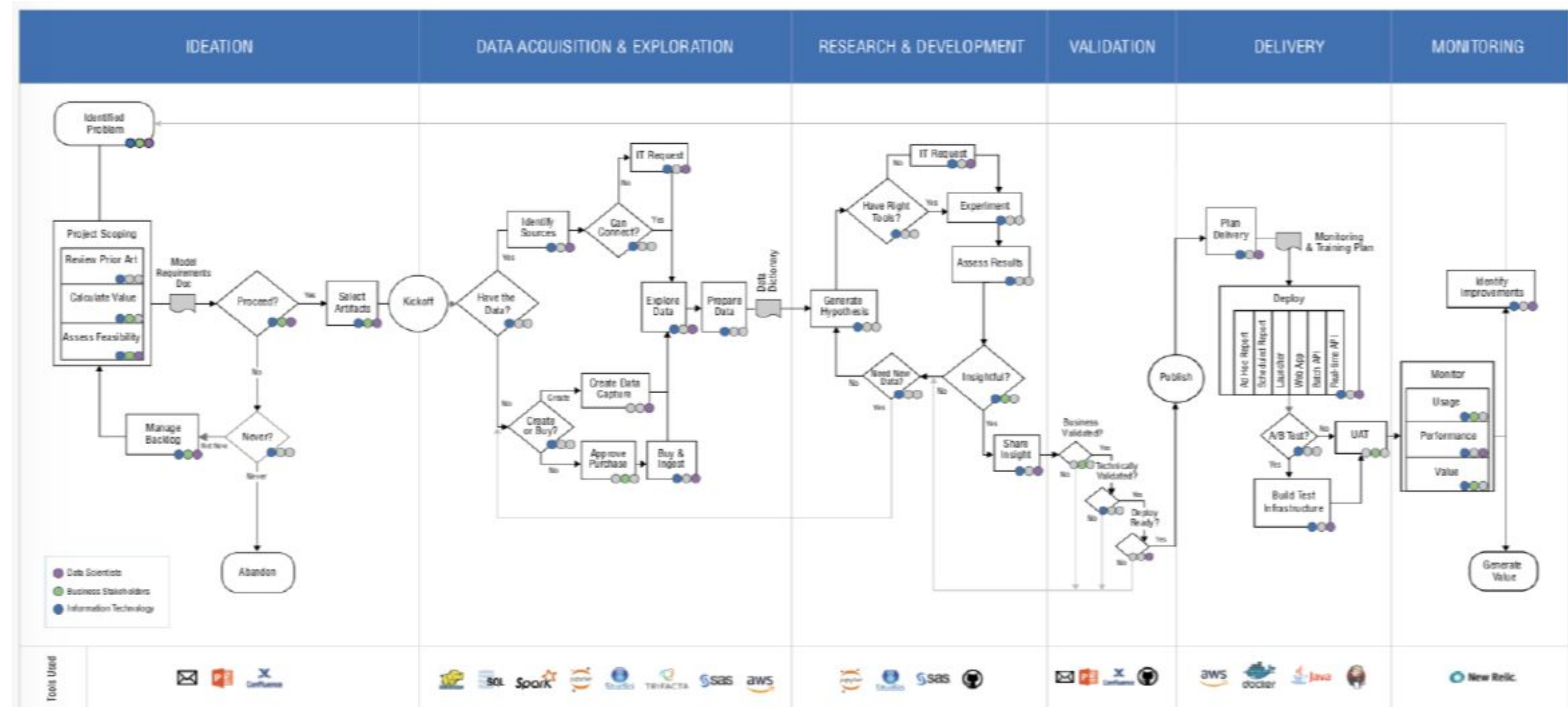
Microsoft Team Data Science  
Process





# Metodologi Lengkap : Kegiatan DS / AI dianggap Kegiatan Bisnis : Masalah Bisnis menjadi masalah DS/AI

Domino DataLab Methodology





# Tahapan Pengembangan



## Business Understanding : Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis ?

### A. Regresi / Estimasi : Memprediksi nilai kontinyu dari kasus

- Prediksi harga rumah berdasar karakteristik tertentu
- Prediksi harga Saham besok

### B. Klasifikasi : Memprediksi kelas/kategori dari kasus

- Prediksi kolektibilitas suatu pinjaman
- Prediksi kebangkrutan perusahaan di masa yang akan datang

### C. Klastering : Mengelompokan kasus berdasar kemiripan

- Segmentasi nasabah perbankan
- Pengelompokan pasien yang mirip kasusnya

### D. Asosiasi : Memprediksi kumpulan item/kejadian yang biasa terjadi Bersama

- Mencari barang jualan yang biasa dibeli Bersama
- Menyusun portofolio saham

### E. Anomali Detection : Menemukan kasus abnormal/tidak biasa terjadi

- Pendeteksian transaksi illegal penggunaan kartu kredit
- Pendeteksian penerobosan network

### F. Sequence Mining : Memprediksi apa yang akan terjadi dari keadaan saat ini

- Prediksi apakah nasabah akan berhenti berlangganan
- Menentukan alur pada transaksi e-commerce

### G. Rekomendasi: Memberikan rekomendasi penggunaan berdasar asosiasi preferensi dengan pengguna lain yang memiliki 'taste' yang sama

- Rekomendasi film untuk ditonton
- Rekomendasi saham untuk dibeli



# Business Understanding : Menentukan Tugas Analytics

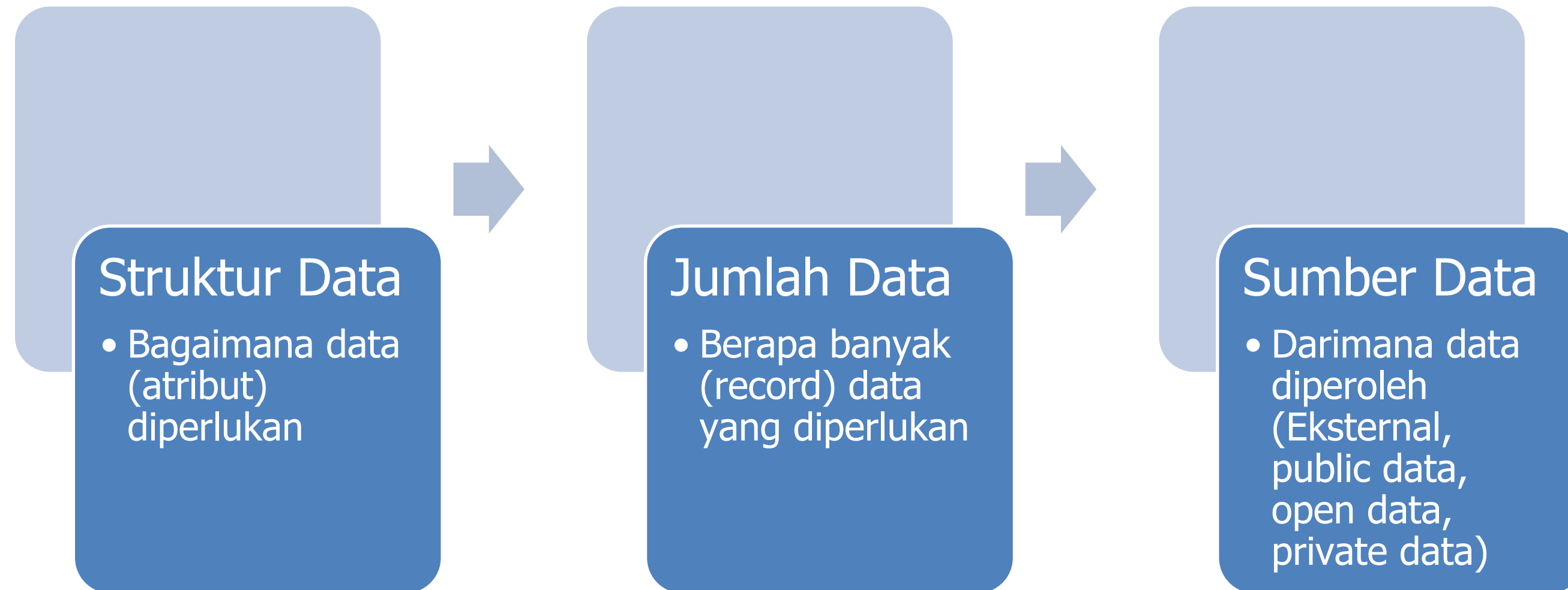
Pengukuran Performansi tergantung jenis TASK ANALYTICS

Metriks Performansi : Ukuran keberhasilan dari proses data science yang dilakukan

Contoh : Root Mean Squared Error (RMSE)

- R-Square
- Jackard Index
- Log-Loss
- Precision
- Recall
- F1-Score

# Business Understanding : Data Apa yang diperlukan ? Dari mana bisa diperoleh ?



# Business Understanding : Merencanakan Manajemen Proyek

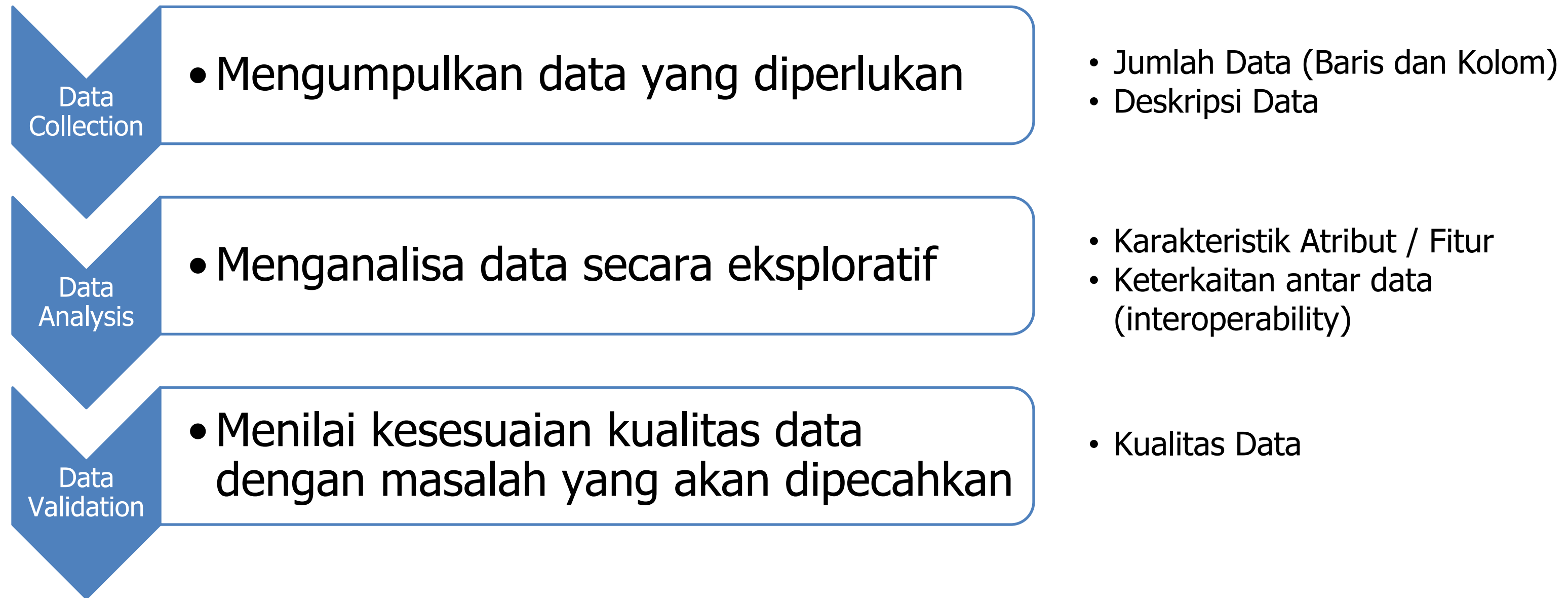
- Bagaimana rencana pelaksanaan proyeknya ?

Cost Benefit Analysis	Situation Assesment	Project Plan
<ul style="list-style-type: none"> <li>• Apakah menguntungkan untuk melakukannya ?</li> </ul>	<ul style="list-style-type: none"> <li>• Analisa keadaan organisasi</li> </ul>	<ul style="list-style-type: none"> <li>• Scope</li> <li>• Time-Schedule</li> <li>• Tim Pengembang</li> </ul>

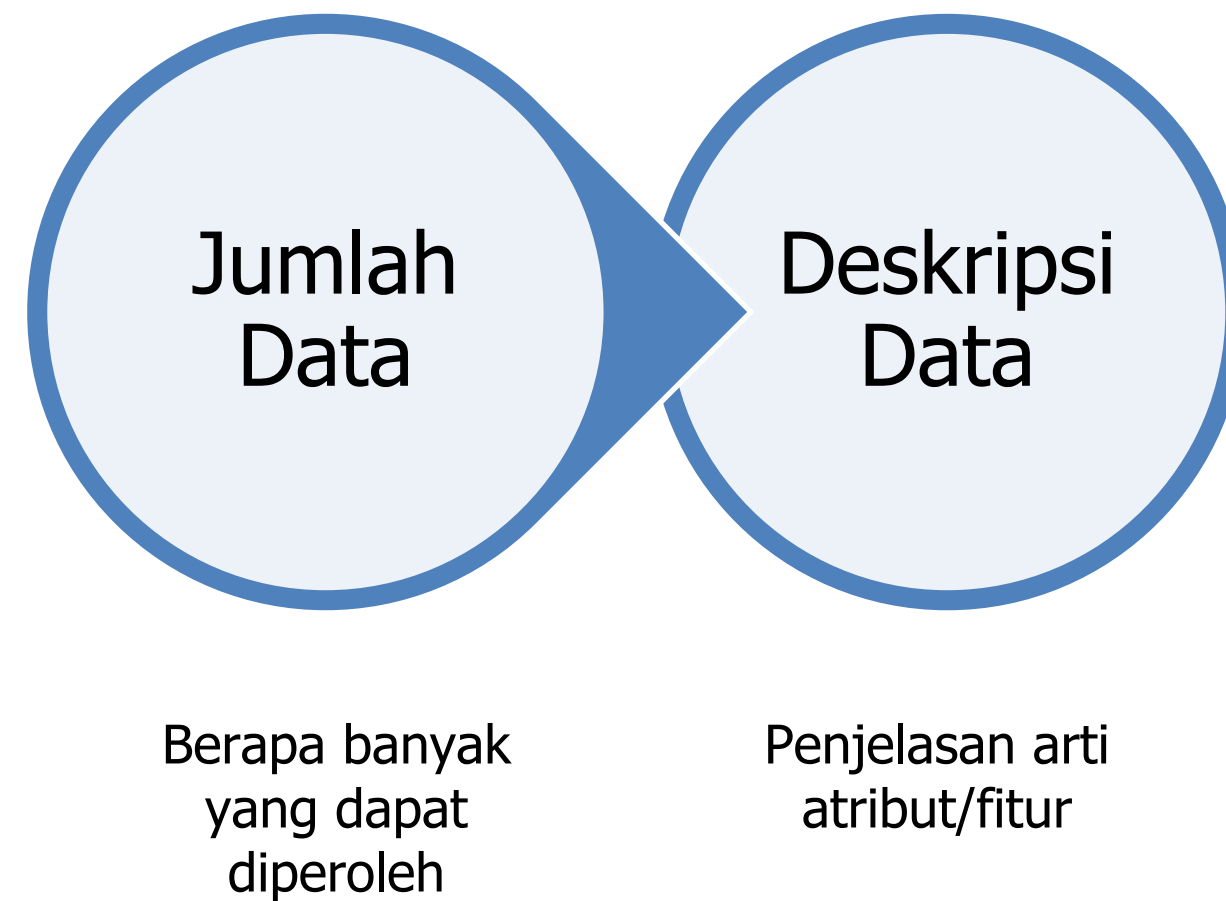


# Data Understanding :

## Mengenal/mendalami data yang dimiliki



# Data Understanding: Mengumpulkan Data



# Data Understanding : Menelaah Data

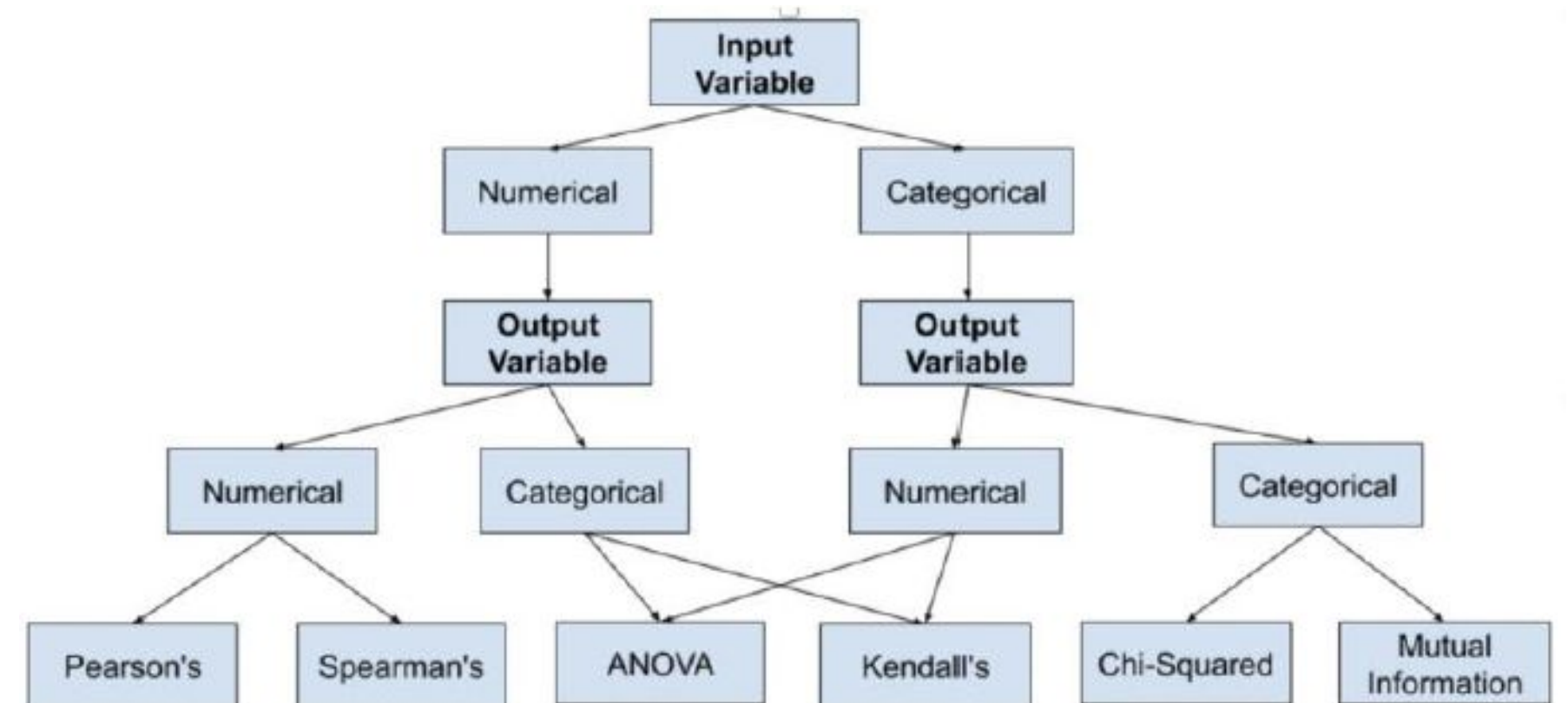
Data ditelaah secara eksploratif (Eksploratif Data Analytic)

## Karakteristik Atribut

- Deskripsi Data (atribut) yang diperoleh

## Keterkaitan Antar Data

- Analisis statistik korelasi, Anova, Chi-Squard,...





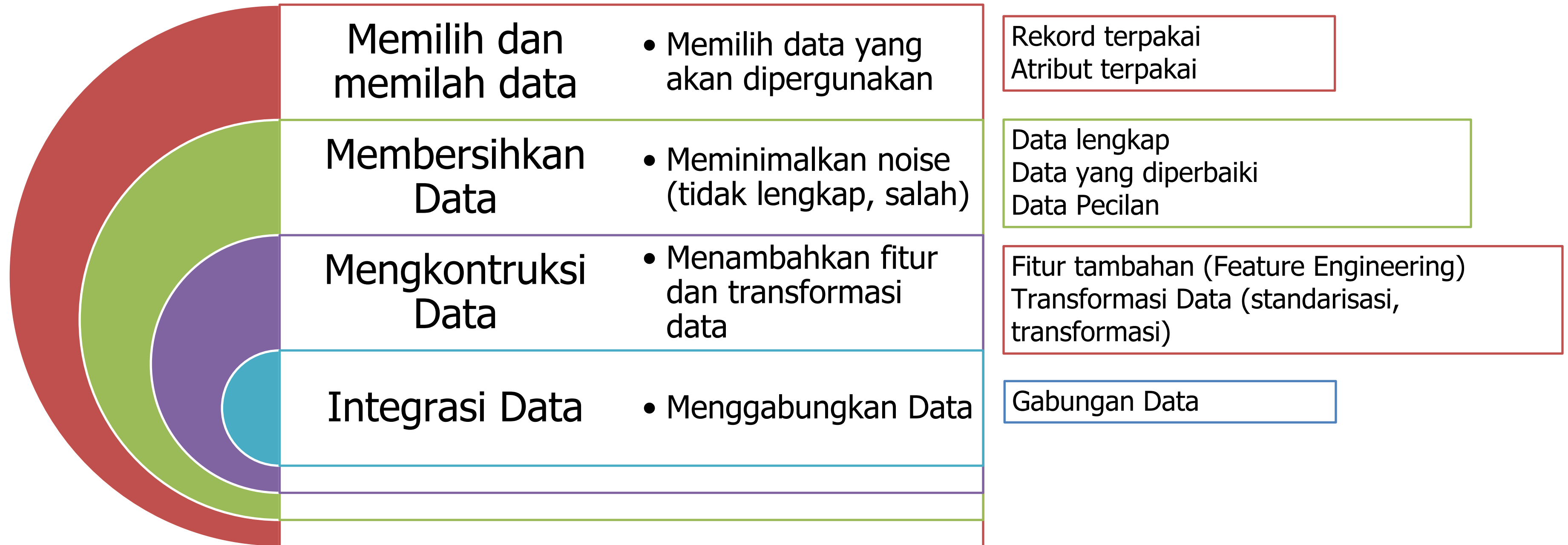
# Data Understanding : Memvalidasi Data

Menilai kesesuaian kualitas data dengan masalah yang dipecahkan

## Laporan Kualitas Data

- Ukuran Data (Atribut/fitur dan Jumlah record)
- Deskripsi statistical atribut
- Relasi antar atribut
- Visualisasi Data

# Data Preparation: Memperbaiki kualitas data untuk Pemodelan



# Modeling: Mengembangkan Mode (Pengetahuan)

## Membangun Skenario Pemodelan

- Membuat strategi pencarian model terbaik

- Pemilihan Algoritma Machine Learning (ML)
- Pembagian Data
- Penentuan Langkah Eksperimen

## Membangun Model

- Mengembangkan Model dengan Teknik ML

- Eksekusi Algoritma
- Pengaturan Parameter
- Pengukuran Performance Metrics

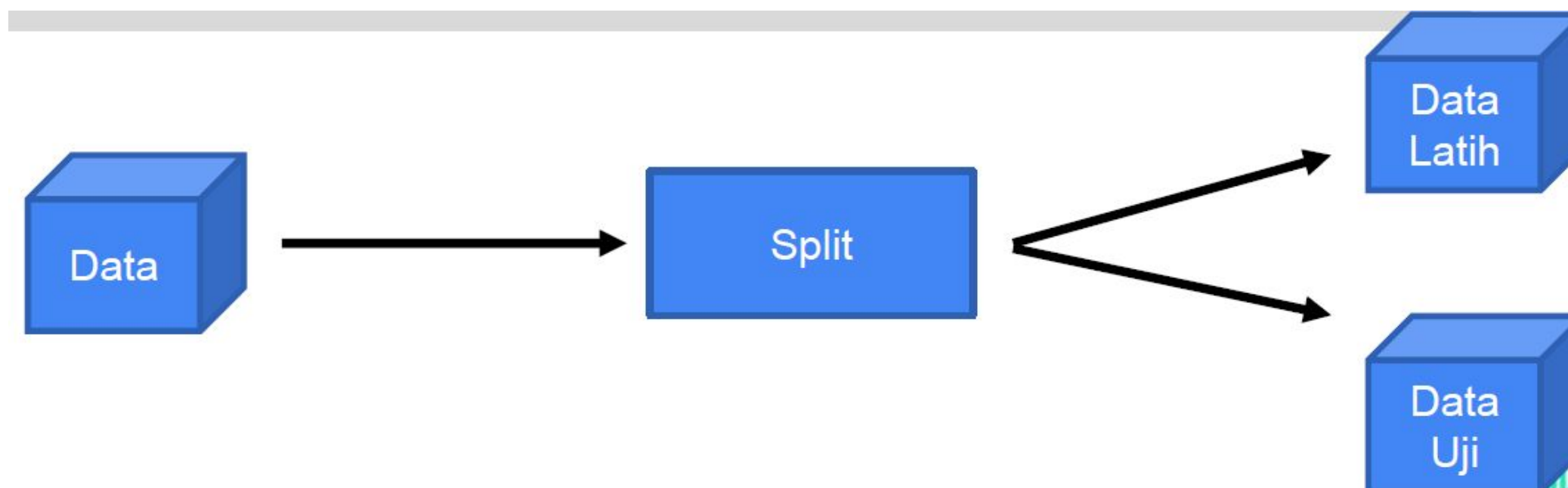


# Modeling :

## Membangun Skenario Pemodelan

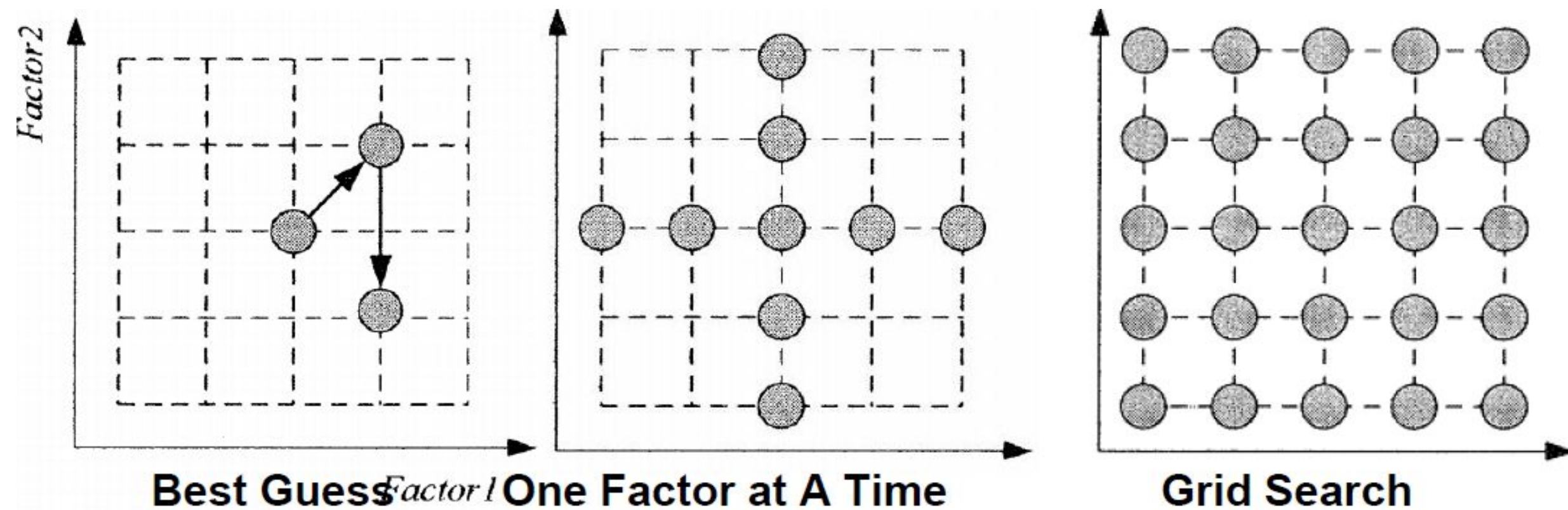
- Memilih Algoritma (d disesuaikan dengan Task Analytic yang dipilih)
  1. K-Nearest Neighbor (k-NN)
  2. Naïve Bayes
  3. Regression Techniques
  4. Support Vector Machine (SVM)
  5. Decision Trees
  6. Random Forest
  7. Deep Learning

- Membagi Data (sesuai dengan ketersediaan data)
  1. Data Latih (Training), untuk mengembangkan model
  2. Data Uji (Testing), untuk mengukur performansi model



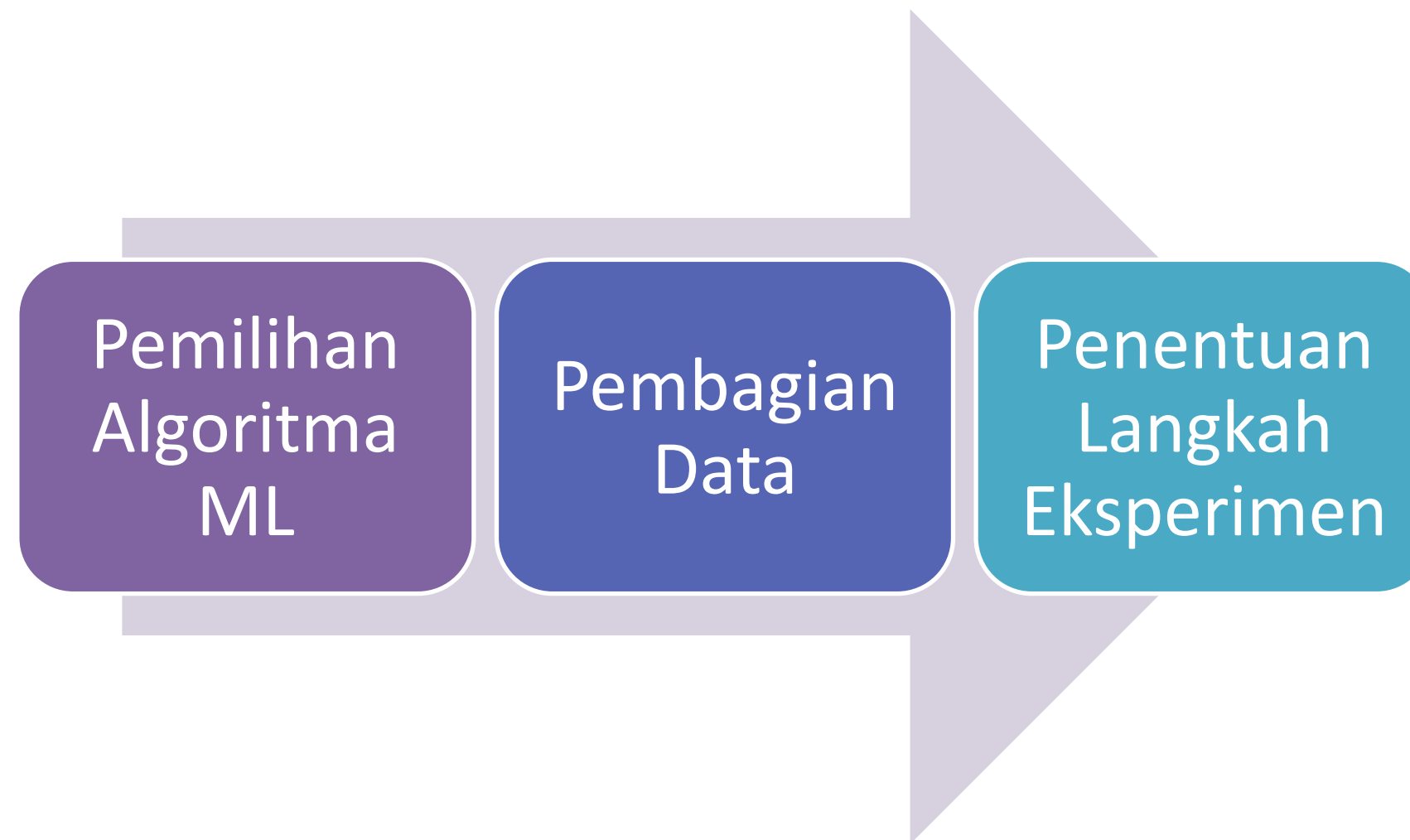
# Modeling : Membangun Skenario Pemodelan

- Menentukan Langkah Eksperimen : Untuk mendapatkan model terbaik secara efisien dan efektif



# Modeling: Membangun Model

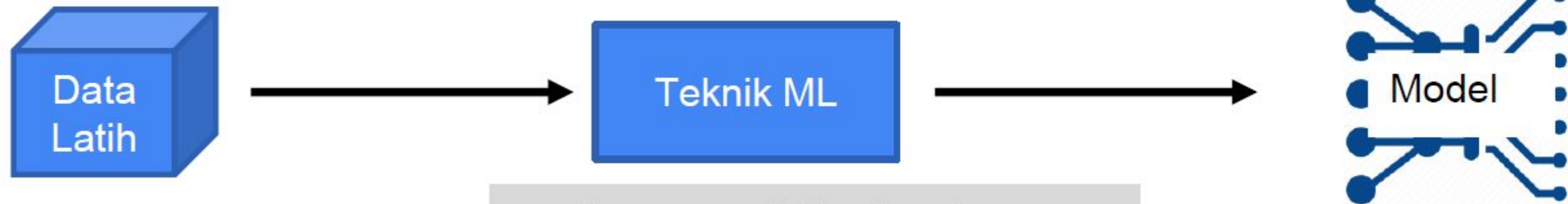
- Mengembangkan model dengan Teknik ML





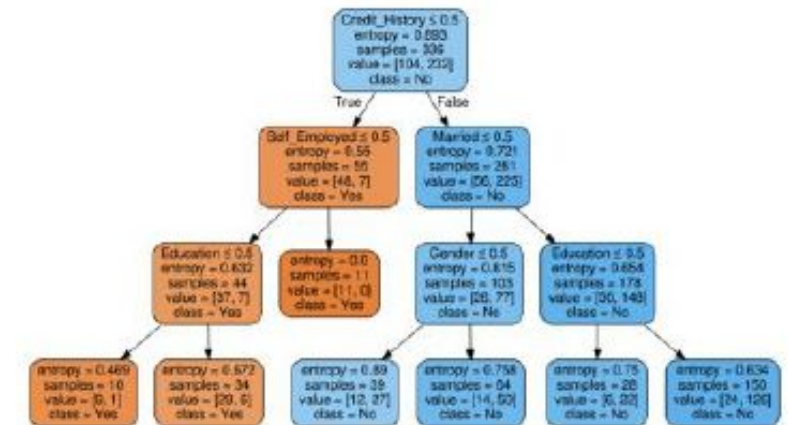
# Modeling : Membangun Model

- Proses Pelatihan : Untuk mendapatkan Model



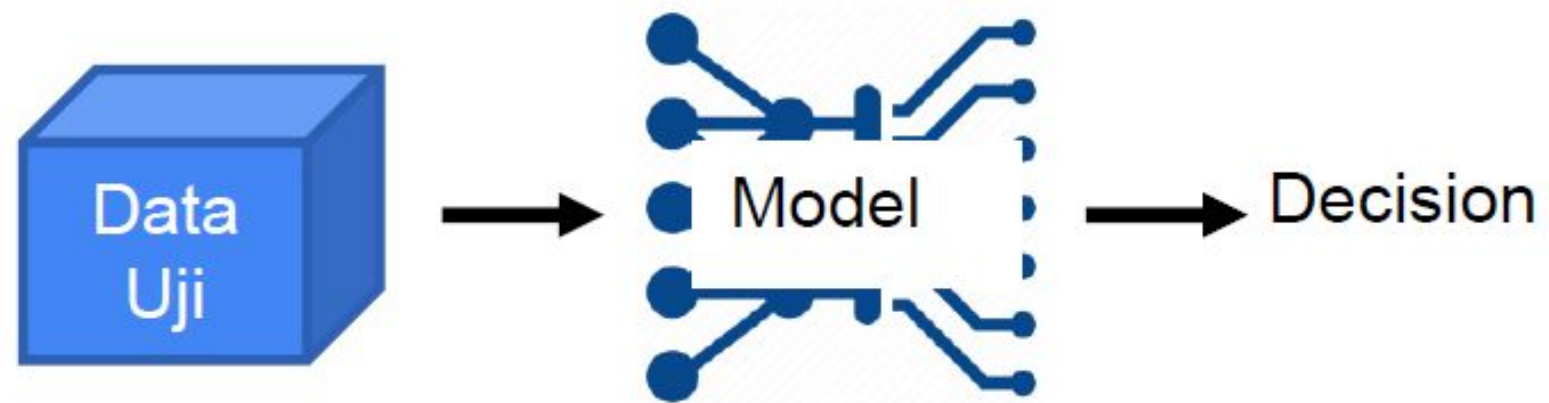
Variable	Type	Definition
BAD	Num	BAD: 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan
LOAN	Num	LOAN: Amount of the loan request
MORTDUE	Num	MORTDUE: Amount due on existing mortgage
VALUE	Num	VALUE: Value of current property
REASON	Char	REASON: DebtCon = debt consolidation; HomeImp = home improvement
JOB	Char	JOB: Occupational categories
YOJ	Num	YOJ: Years at present job
DEROG	Num	DEROG: Number of major derogatory reports
DELINQ	Num	DELINQ: Number of delinquent credit lines
CLAGE	Num	CLAGE: Age of oldest credit line in months
NIINQ	Num	NIINQ: Number of recent credit inquiries
CLNO	Num	CLNO: Number of credit lines
DEBTINC	Num	DEBTINC: Debt-to-income ratio

1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. ...



# Modeling: Membangun Model

- Proses Pengujian: Untuk mengukur Performansi



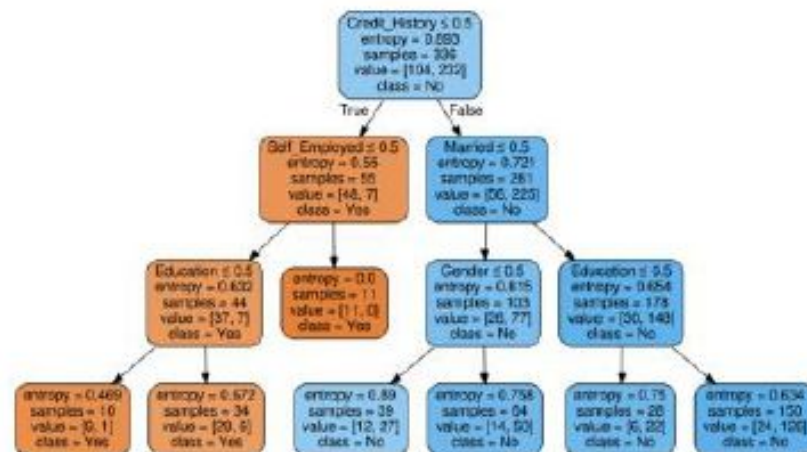
**TP** = True Positives  
**TN** = True Negatives  
**FP** = False Positives  
**FN** = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



# Model Evaluation

- Mengevaluasi Performance Model yang dihasilkan

## Mengevaluasi Model

- Mengukur performance Model

Performansi Capaian vs Target  
Memilih Model Terbaik

## Mengevaluasi Proses

- Menilai apakah proses sudah maksimal

Review Proses untuk mencari  
Batasan atau kekurangan model



# Prospek Kerja dan Profesi Data Scientist Di Masa Depan

## Are these the world's best jobs?

Ranking determined by work-life balance rating

Rank	Job	Salary
1	Data Scientist	\$114,808
2	SEO Manager	\$45,720
3	Talent Acquisition Specialist	\$63,504
4	Social Media Manager	\$40,000
5	Substitute Teacher	\$24,380
6	Recruiting Coordinator	\$44,700
7	UX Designer	\$91,440
8	Digital Marketing Manager	\$70,052
9	Marketing Assistant	\$32,512
10	Web Developer	\$66,040
11	Risk Analyst	\$69,088
12	Civil Engineer	\$65,532
13	Client Manager	\$71,120
14	Instructional Designer	\$66,040
15	Marketing Analyst	\$60,000
16	Software QA Engineer	\$91,440
17	Web Designer	\$53,848
18	Research Technician	\$36,525
19	Program Analyst	\$71,120
20	Data Analyst	\$58,928
21	Content Manager	\$60,960
22	Solutions Engineer	\$92,456
23	Lab Assistant	\$27,550
24	Software Developer	\$80,000
25	Front End Developer	\$75,000

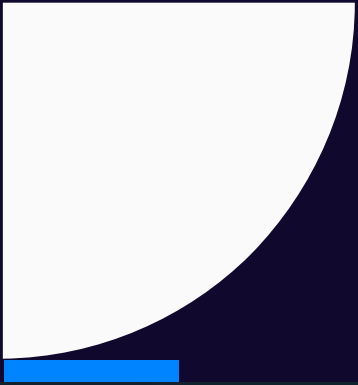


# Contoh Industri Yang Memerlukan Data Scientist





# Perbedaan DS, DA dan DE



**Data Scientist**



**Data Analyst**



**Data Engineer**

**V S V S**



# Perbedaan DS, DA dan DE

## Data Scientist



Building model using Machine Learning, handle very specific cases or problems

## Data Engineer



Build and maintain data structures and architectures for data ingestion, processing, and deployment for large-scale data intensive applications

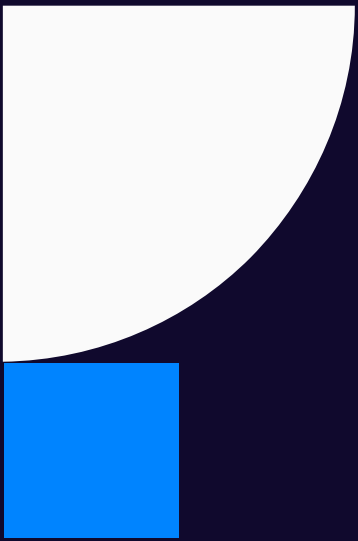
## Data Analyst



Dissect and diagnose product/ business problem using data



# Perbedaan DS, DA dan DE



## Skillset

### Data Scientist

#### Programming Languages

Python, R, SQL, SAS, Java

#### Frameworks

Pig, Spark, Hadoop

#### Technologies

Machine Learning, Deep Learning

### Data Analyst

#### Programming Languages

Python, R, SQL, SAS, JavaScript

#### Tools

SAS Miner, Microsoft Excel, SSAS, SPSS

### Data Engineer

#### Programming Languages

Python, R, SQL, SAS, Java

#### Frameworks

Hadoop, MapReduce, Hive, Pig, Apache Spark, Data Streaming, NoSQL



# Rangkuman

- Langkah-Langkah utama dalam menggunakan data untuk membuat suatu aplikasi AI berdasarkan metodologi data science
- Pengembangan Sistem AI berdasarkan data bukan hanya masalah teknis (terkait data) namun merupakan masalah bisnis/organisasi
- Pengembangan sistem melibatkan Pakar Domain, Pakar Data Science/AI, Pakar Manajemen Project dan Pakar TI dalam satu Tim

# Referensi

- CRISP-DM  
<http://crisp-dm.eu/>
- IBM Data Science Methodology  
<https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science>
- Microsoft Methodology  
<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>
- Domino Methodology  
<https://www.dominodatalab.com/>