Data Science Methodology

Profile

Mathematics

Data Scientist –
Stream Intelligence

Senior Data Analyst –
Tokopedia

Farhan Reza Gumay

Quote Related Today

*"In recent years, all the headlines about big data, business intelligence, analytics, performance management, data lakes, and AI have diverted our attention from the ultimate reason for these solutions – **decision support**."*

Dan Vesset

Group Vice President of the Analytics and Information Management Market Research and Advisory Practice at IDC

Table of Content

What will We Learn Today?

1. Understand data analytics lifecycle
2. Explore the common methodology for data science

Data Analytics Lifecycle

This approach drives business value and innovation through continuous improvement

Process Details

Business Understanding

What's your plan?

Planning relies on outputs of all othersteps.

- It requires an understanding of past performance
- Identification of deviations from the norm (plan vs. actual)
- Evaluation of possible scenarios
- Prediction of likely outcomes, and assessment of risks and constraints

Process Details

Performance Analysis

What happened?

Associated with data visualization via reports, dashboards, and scorecards that facilitates decision makings

- State business metrics
- Identify data required
- Extract and prepare data
- Analyze data
- Present data

Process Details

Identify Business Causes

Why did it happen?

Uncover details such as the frequency of events, the cost of operations and the root cause of failures

- Identify anomalies
- Drill into the analytics (discovery)
- Determine causal relationships

Process Details

Predict Using Data Models

What happens next?

Using descriptive data accumulated over time, predictive analytics utilizes models for predicting events

- Identify business outcomes
- Determine data required to train
- Determine types of analysis
- Validate results
- Test predictions on performance

Process Details

Solution Optimization

What should you do?

Use advance capabilities such us optimization and mathematical models to reveal not only recommended actions, but also why they are recommended along with any implications the actions might have

- Input data from model prediction
- Combine data models with rules
- Provide constraint-based optimization
- Implement for better decisions

Data Science Methodology

Provides the data scientist with a framework for how to proceed with whatever methods, processes, and heuristics will be used to obtain answers or result

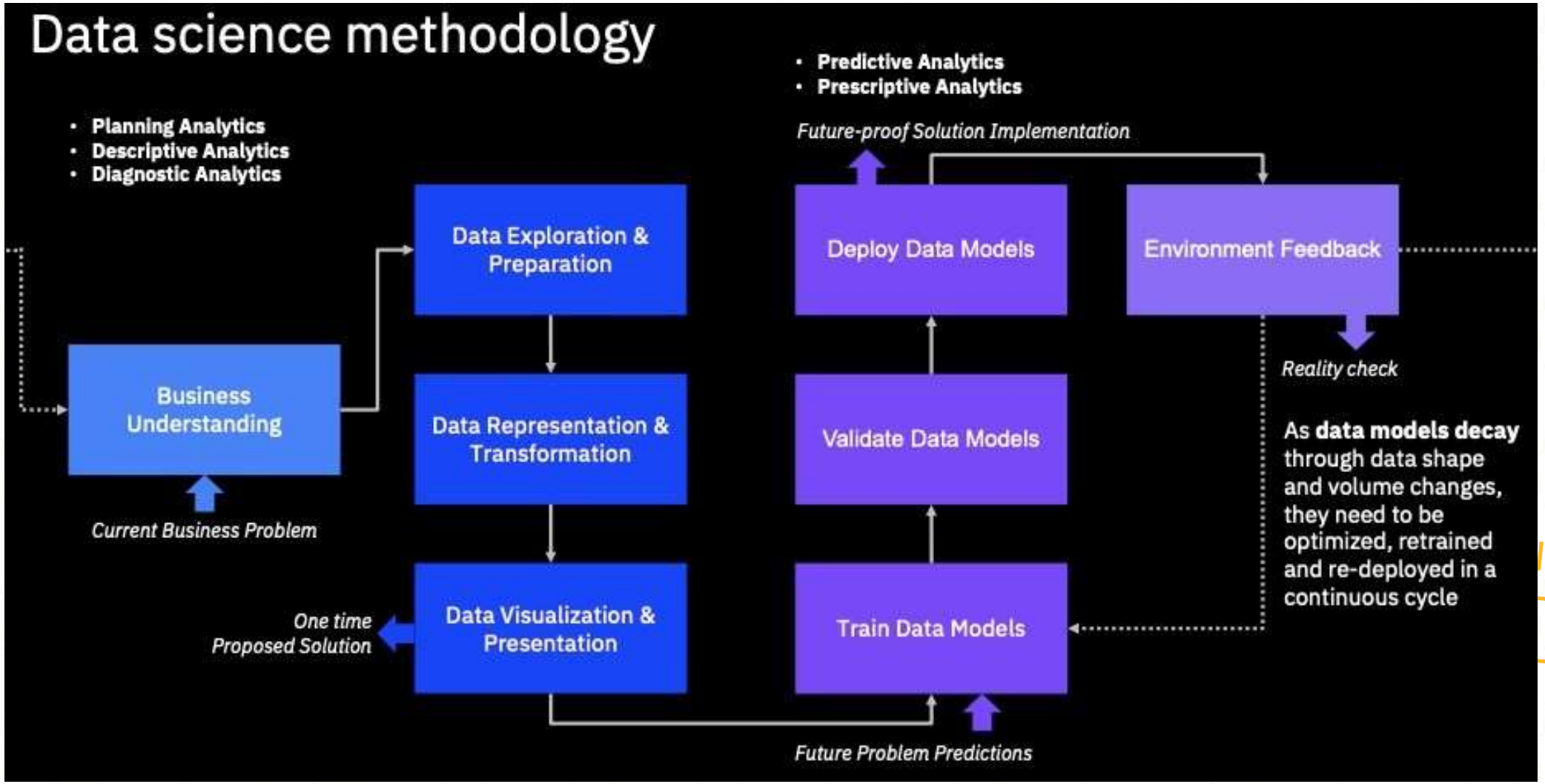
The Need for a Data Science Methodology

Data scientists need a foundational methodology that also addresses **unstructured data** and **prescriptive analytics**

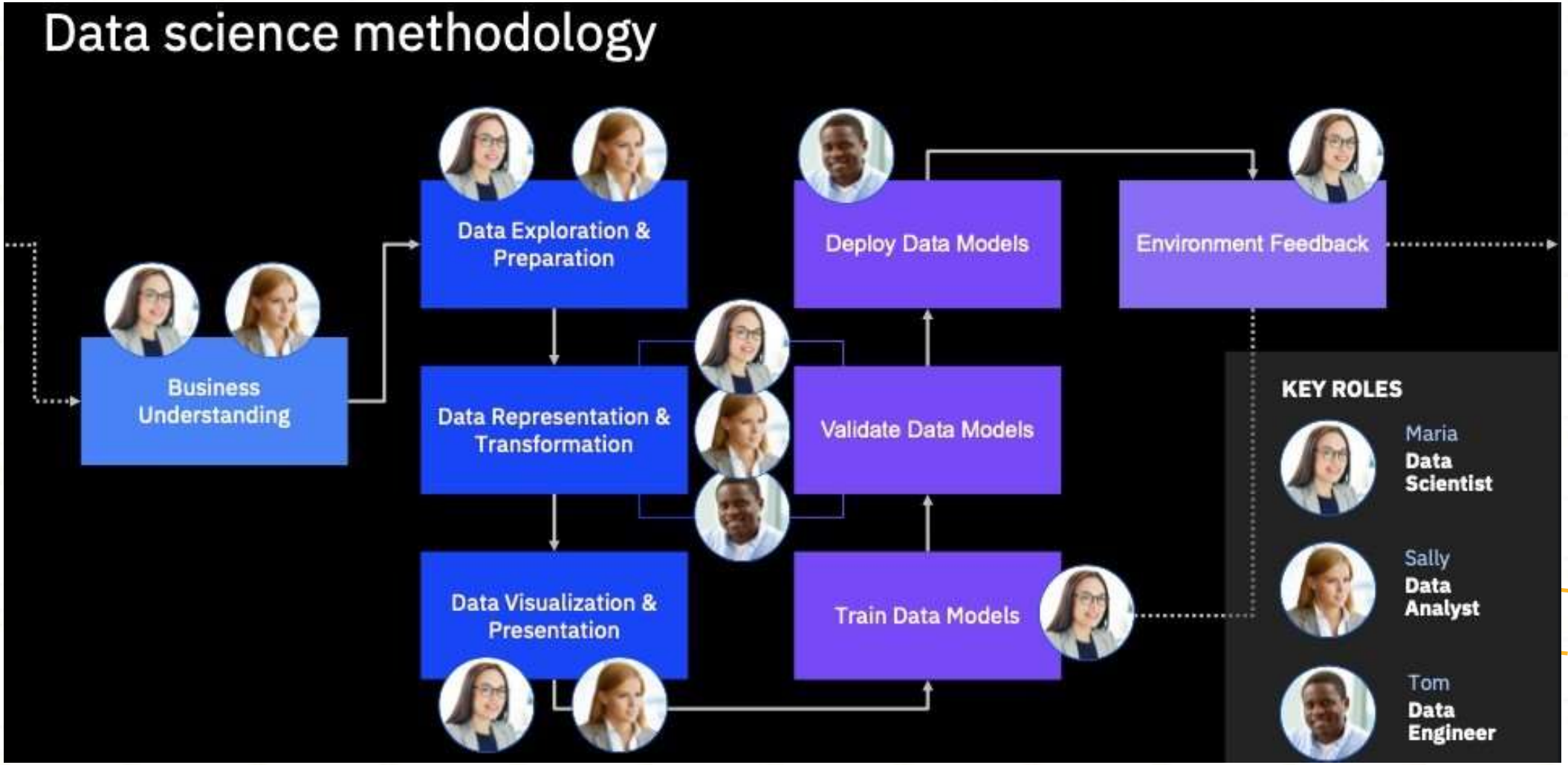
Emphasizing on new practices in data science such as:

- Use of **very large volumes of data**
- Incorporation of **Machine Learning capabilities** into **prescriptive modelling and automation**

Data Science Methodology



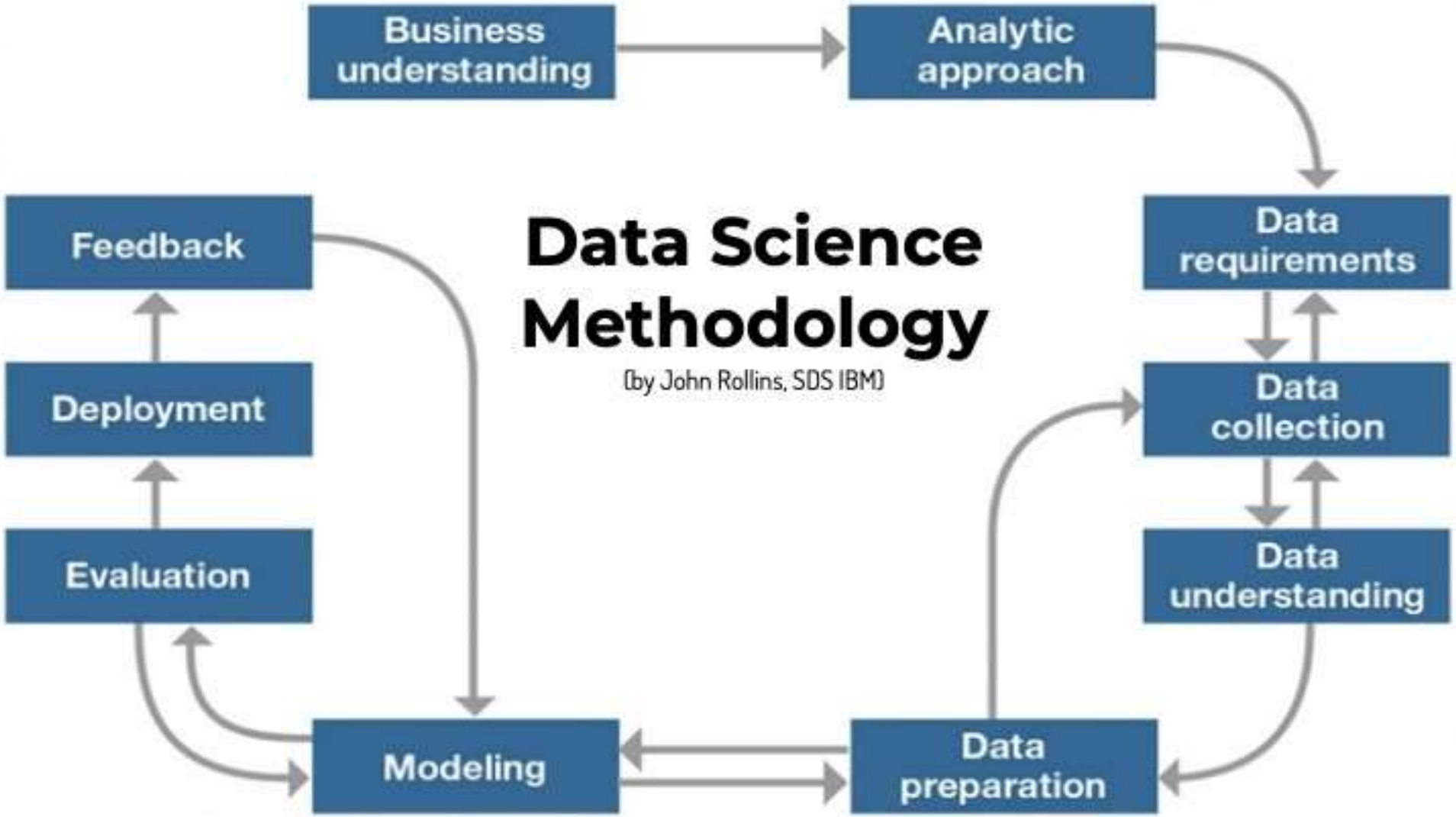
Data Science Methodology



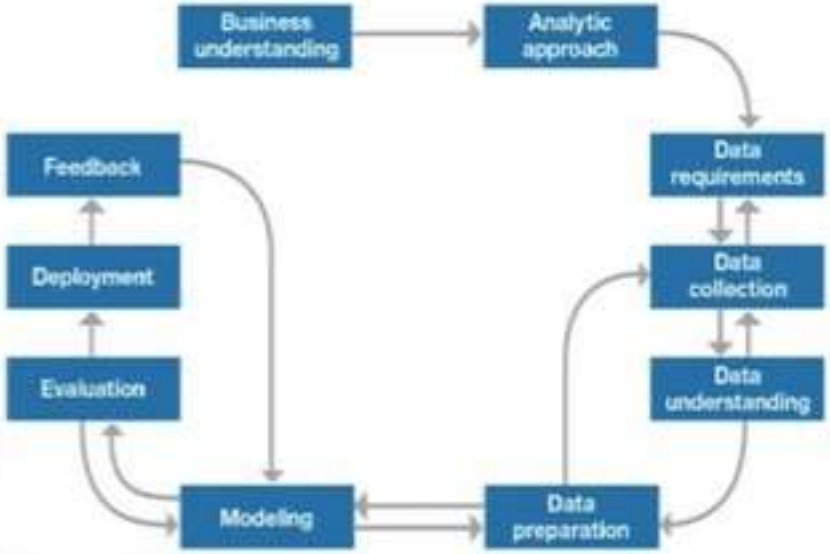
Data Science Methodology

Data Science Methodology's purpose is **to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated** to address the question at hand.

The data science methodology discussed in this course has been outlined by John Rollins, a seasoned and senior data scientist currently practicing at IBM.



In a nutshell…



The **Data Science Methodology** aims to answer the following 10 questions in this prescribed sequence:

From problem to approach:

1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

Working with the data:

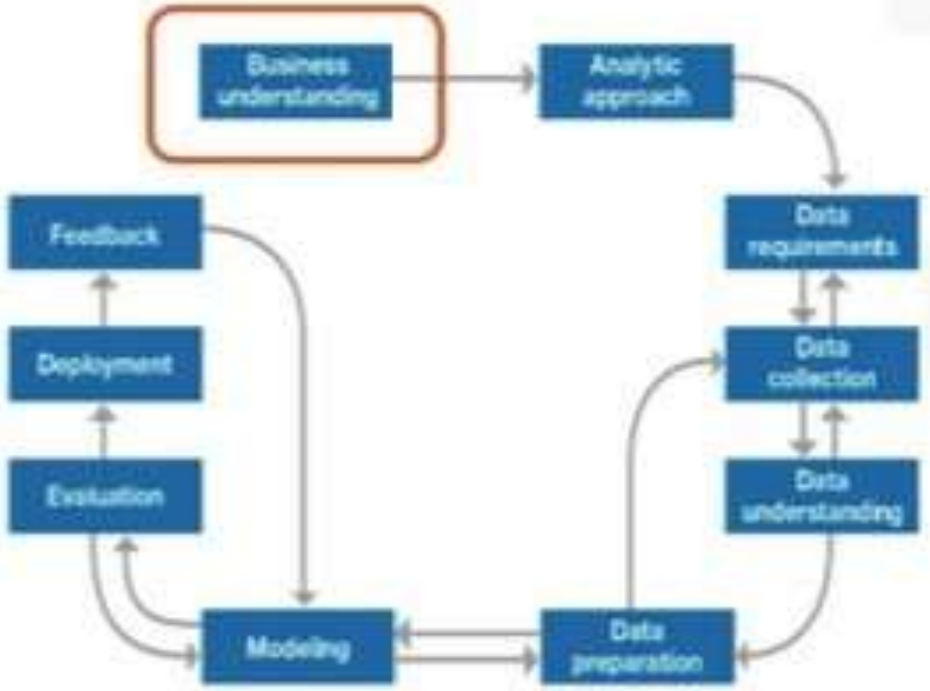
3. *What data do you need to answer the question?*
4. *Where is the data coming from (identify all sources) and how will you get it?*
5. *Is the data that you collected representative of the problem to be solved?*
6. *What additional work is required to manipulate and work with the data?*

Deriving the answer:

7. *In what way can the data be visualized to get to the answer that is required?*
8. *Does the model used really answer the initial question or does it need to be adjusted?*
9. *Can you put the model into practice?*
10. *Can you get constructive feedback into answering the question?*

Business

Understanding



Business Understanding

Business understanding

- *What is the problem that you are trying to solve?*

Business Understanding

Rollins suggests that **having a clearly defined question is vital because it ultimately directs** the analytic approach that will be needed to address the question.

All too often, much effort is **put into answering what people THINK is the question, and while the methods used to address that question might be sound**, they don't help to solve the actual problem.

Establishing a clearly defined question starts with understanding the GOAL of the person who is asking the question.

Business Understanding

The key business sponsors involvement throughout the project was critical, in that the sponsor:

- Set overall direction
- Remained engaged and provided guidance
- Ensured necessary support, where needed

Understand the Business Metrics

DigitalSkola
Uncover The World Of Digital Skills With Us

Digital Marketing

visitors
new visitors
app installs
app rating

Transactions

revenue
transactions
unique buyers
Conversion Rate (CvR)

Promo

promo cost
ROI (Return of Investment)

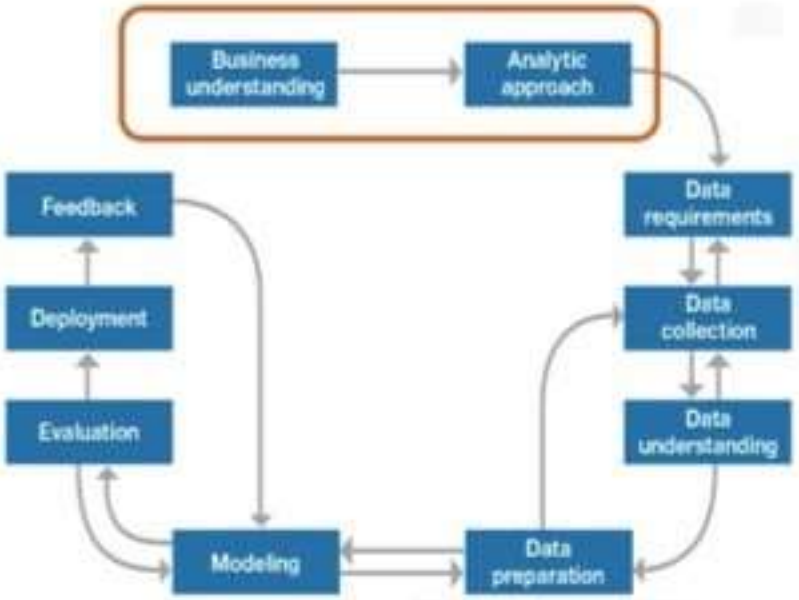
Customer Service

complaints
resolved complaints
CSAT (Cust Satisfaction)

Fraud

caught frauds
cost loss

From Understanding to Approach



Business understanding

- *What is the problem that you are trying to solve?*

Analytic approach

- *How can you use data to answer the question?*

Pick analytic approach based on type of question

Descriptive
- Current status

Diagnostic (Statistical Analysis)
- What happened?
- Why is this happening?

Predictive (Forecasting)
- What if these trends continue?
- What will happen next?

Prescriptive
- How do we solve it?

What are the Types of Questions?

If the question is to determine probabilities of an action
- Use a Predictive model

If the question is to show relationships
- Use a descriptive model

If the question requires a yes/no answer
- Use a classification model

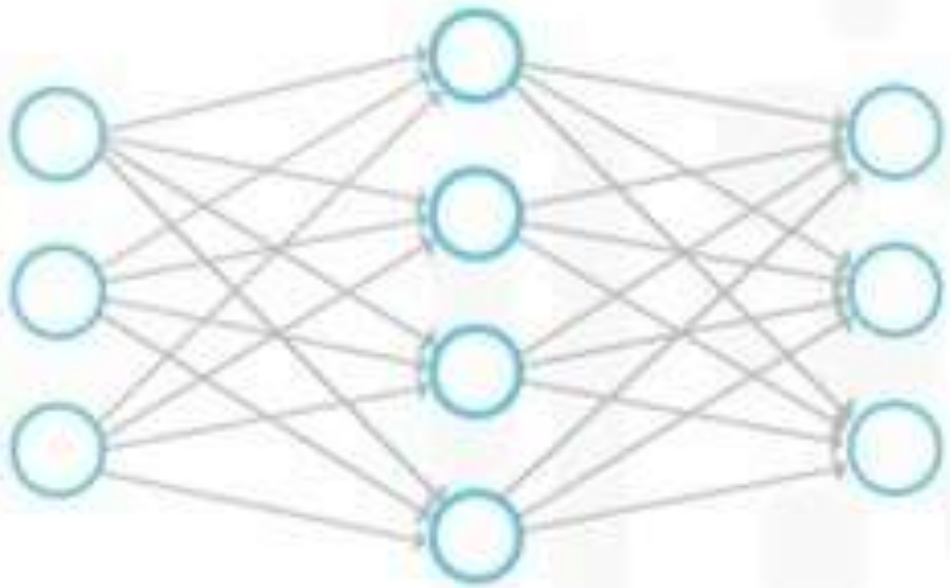
Analytic approach
- *How can you use data to answer the question?*

- The correct approach depends on business requirements for the model

Business Understanding

- Statistical analysis applies to problems that require counts
- **For example if the question requires a yes/ no answer, then a classification approach to predicting a response would be suitable**
- **Machine Learning** is a field of study that gives computers the ability to learn without being explicitly programmed
- Machine Learning can be used to identify relationships and trends in data that might otherwise not be accessible or identified

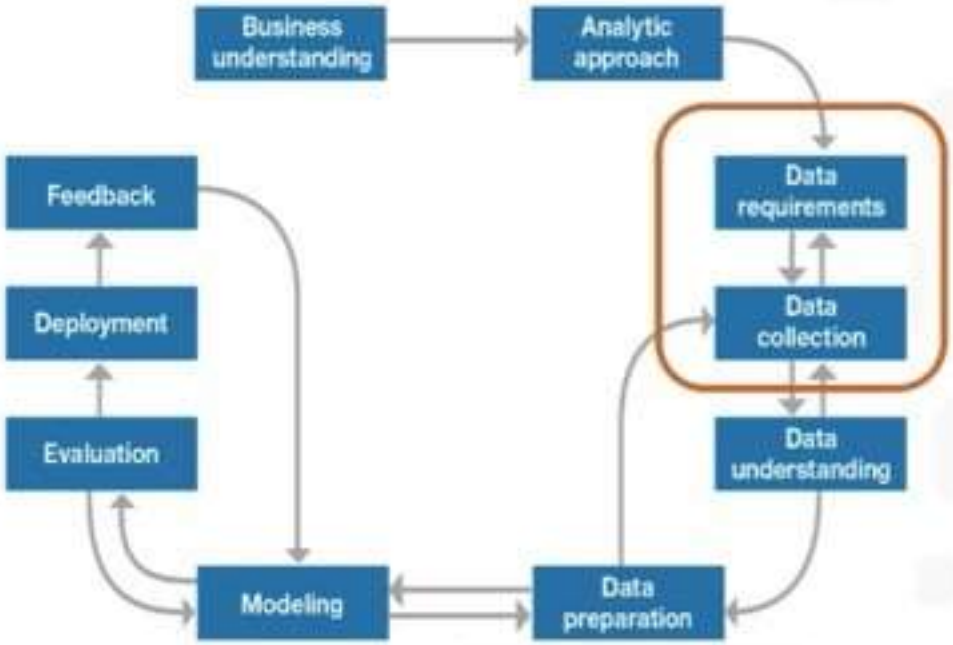
Will Machine Learning will be utilized?



Machine Learning

- Learning without being explicitly programmed
- Identifies relationships and trends in data that might otherwise not be accessible or identified
- Uses clustering association approaches

From Requirements to Collection



Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

Banyak cara mendapatkan data ...

DigitalSkola
Uncover The World Of Digital Skills With Us

Internal

Database Production

data transaksi
data user
data product

Database Events

data user click
data user page view
data user scroll
tracker infrastruktur

Documents

file excel
kumpulan gambar

Eksternal

Data Public

open data
data repository
public dashboard

Data 3rd Party

data dari vendor
data survey

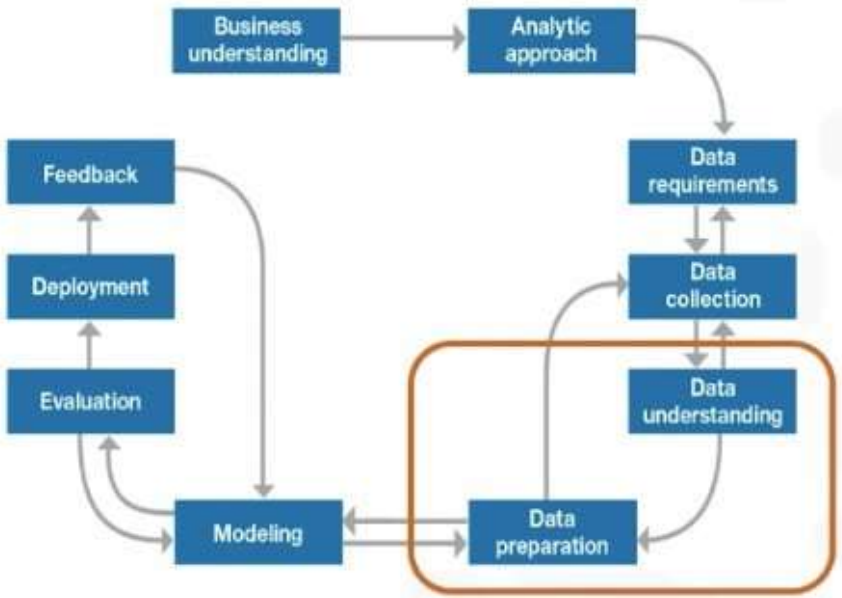
Scraping / API

website
sosial media

Data Collection

- Once the data ingredients are collected, then in the data collection stage, the data scientist will have a good understanding of what they will be working with
- **Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.**
- Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

From Understanding to Preparation



Data understanding

- *What does it mean to "prepare" or "clean" data?*

Data preparation

- *What are ways in which data is prepared?*

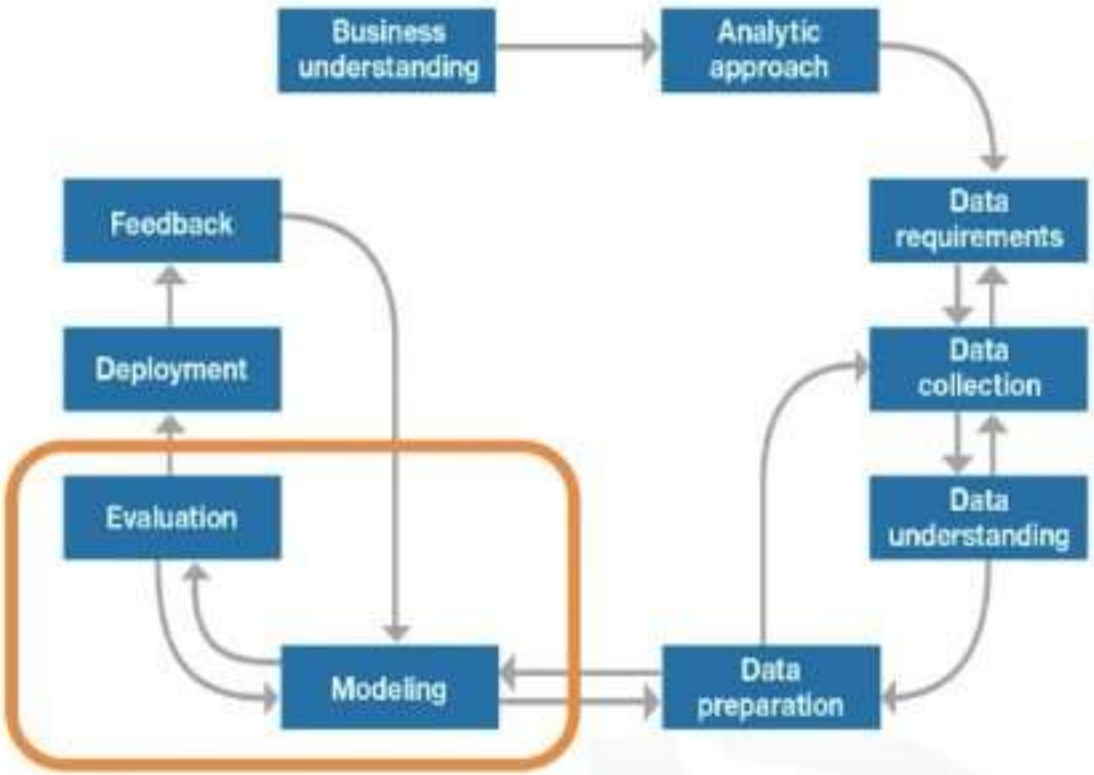
Data Understanding

- Data understanding encompasses **all activities related to constructing the data set**.
- Essentially, the data understanding section of the data science methodology answers the question: **Is the data that you collected representative of the problem to be solved?**
- Statistics needed to be run against the data columns that would become variables in the model.
- The more one works with the problem and the data, the more one learns and therefore the more refinement that can be done within the model, ultimately leading to a better solution to the problem

Data Preparation

1. Handle Missing Data
2. Handle Invalid Values
3. Remove Duplicates
4. Formatting
5. Feature Engineering

From Modeling to Evaluation



Modeling

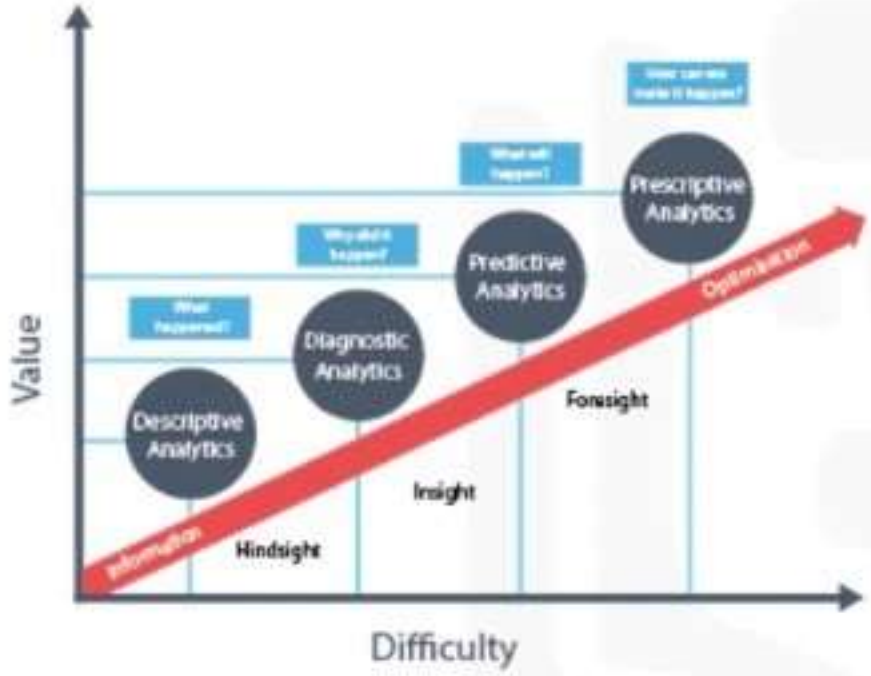
- *In what way can the data be visualized to get to the answer that is required?*

Evaluation

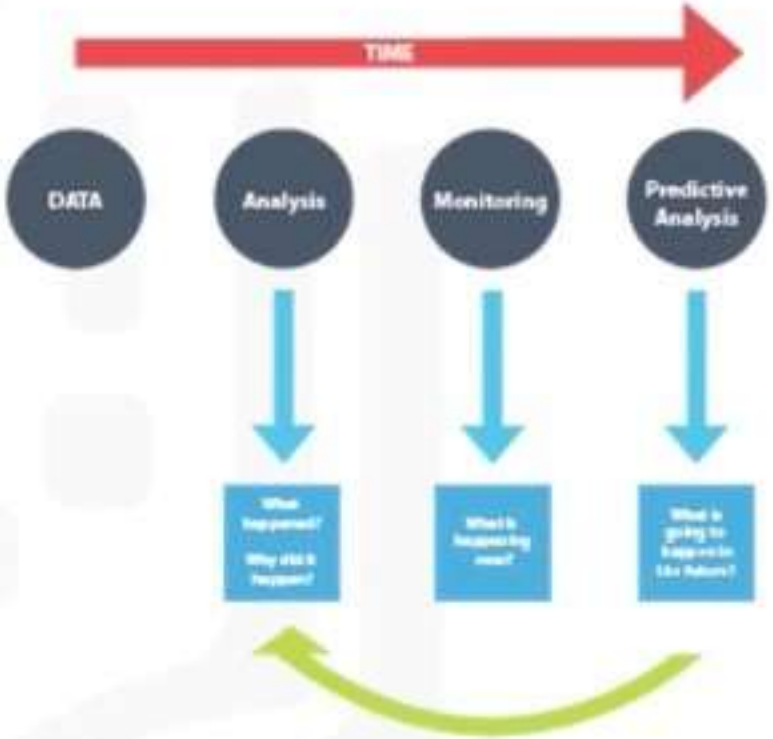
- *Does the model used really answer the initial question or does it need to be adjusted?*

Data Modeling – Using Predictive or Descriptive?

Descriptive Analytics



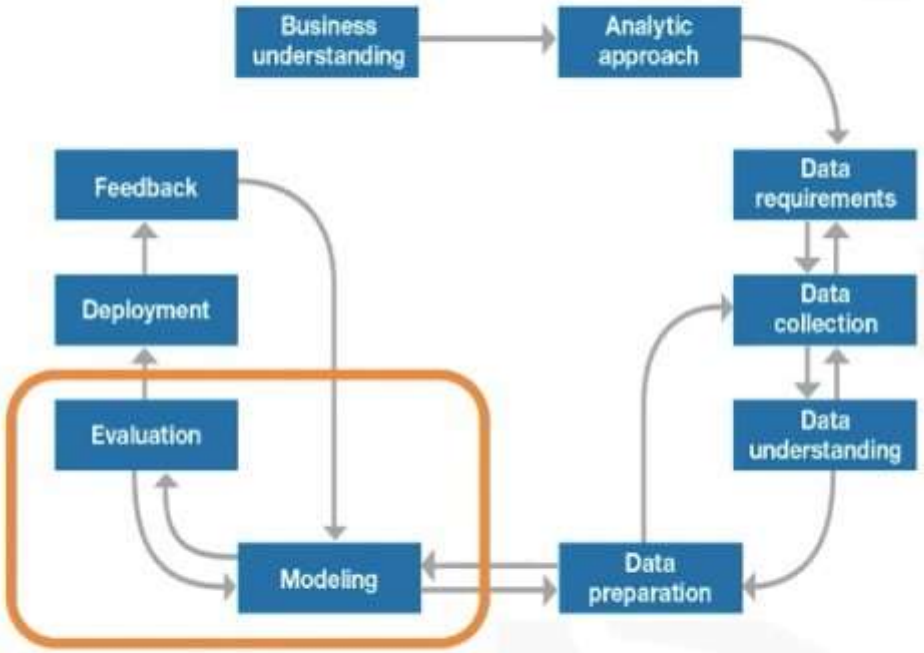
Predictive Analytics



Data Modelling

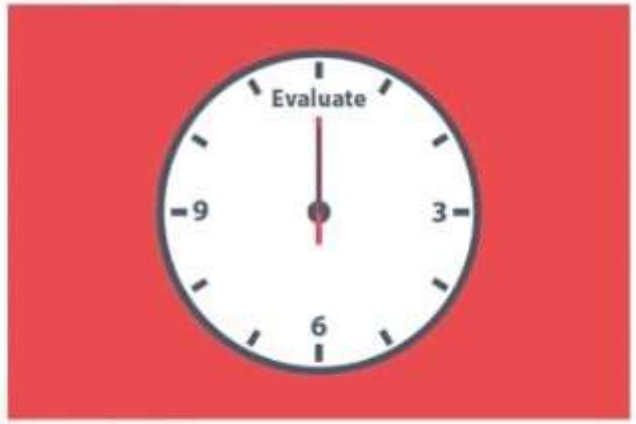
- A predictive model **tries to yield yes/no, or stop/go type outcomes**.
- These models are based on the analytic approach that was taken, either statistically driven or machine learning driven.
- The data scientist will use **a training set** for predictive modelling.
- **A training set is a set of historical data in which the outcomes are already known**.
- The training set acts like a gauge to determine if the model needs to be calibrated.
- In this stage, **the data scientist will play around with different algorithms** to ensure that the variables in play are actually required.

From Modeling to Evaluation



Modeling

- *In what way can the data be visualized to get to the answer that is required?*



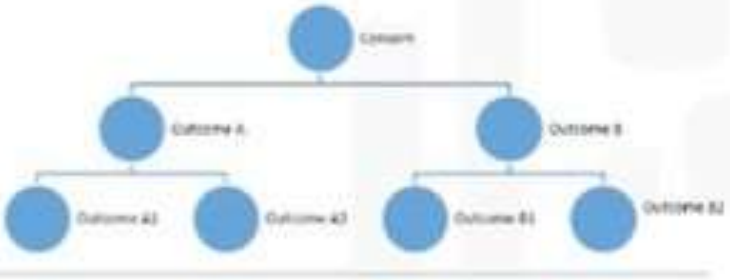
Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

When and how to adjust the model?

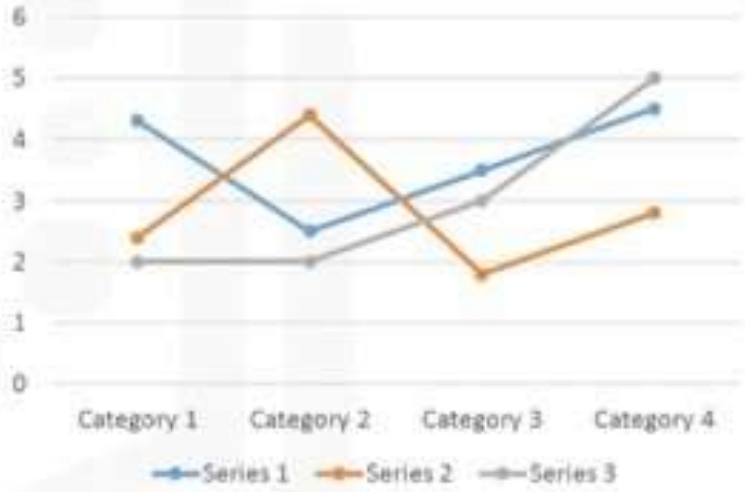
Diagnostic measures

Predictive Model

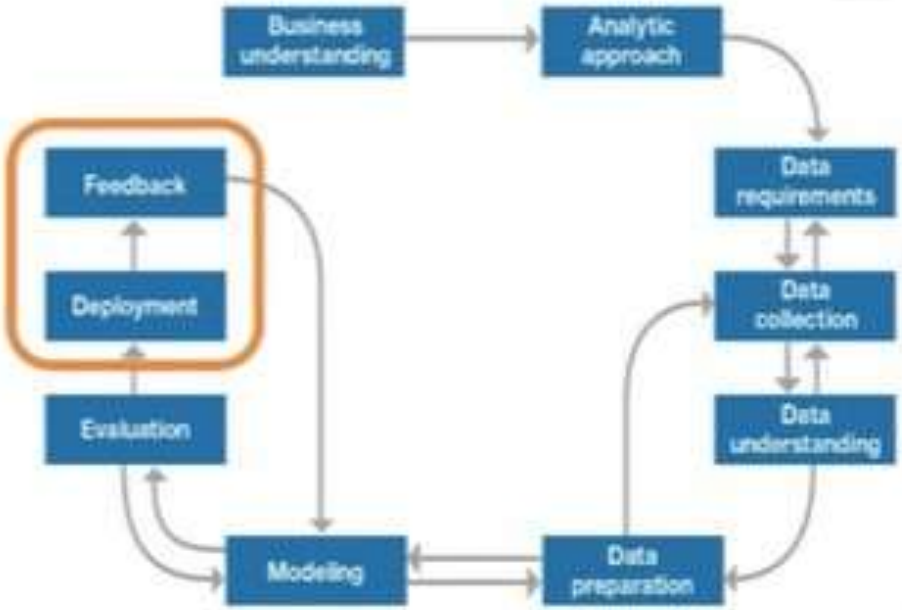


Descriptive Model

Statistical Significance



From Deployment to Feedback



Data Science Methodology artinyaa…..

Langkah-langkah apa saja yang harus dilakukan data scientist untuk menyelesaikan masalah?

10 pertanyaan dasar; 5 aspek; 3 grup:

1. From Problem to Approach
2. From Requirements to Collection
3. From Understanding to Preparation
4. From Modeling to Evaluation
5. From Deployment to Feedback

Perencanaan

Bekerja dengan data

Membangun solusi

[Business Understanding]

1.Problem apa yang ingin kamu selesaikan?

Mulai dengan menentukan **goals** & **objectives**

Goals : *Improvement* apa yang mau dicapai?

Objectives : Apa yang perlu dilakukan untuk mencapai goals tersebut?

Pahami business metrics!

DigitalSkola
Uncover The World Of Digital Skills With Us

Digital Marketing

visitors
new visitors
app installs
app rating

Transactions

revenue
transactions
unique buyers
Conversion Rate (CvR)

Promo

promo cost
ROI (Return of Investment)

Customer Service

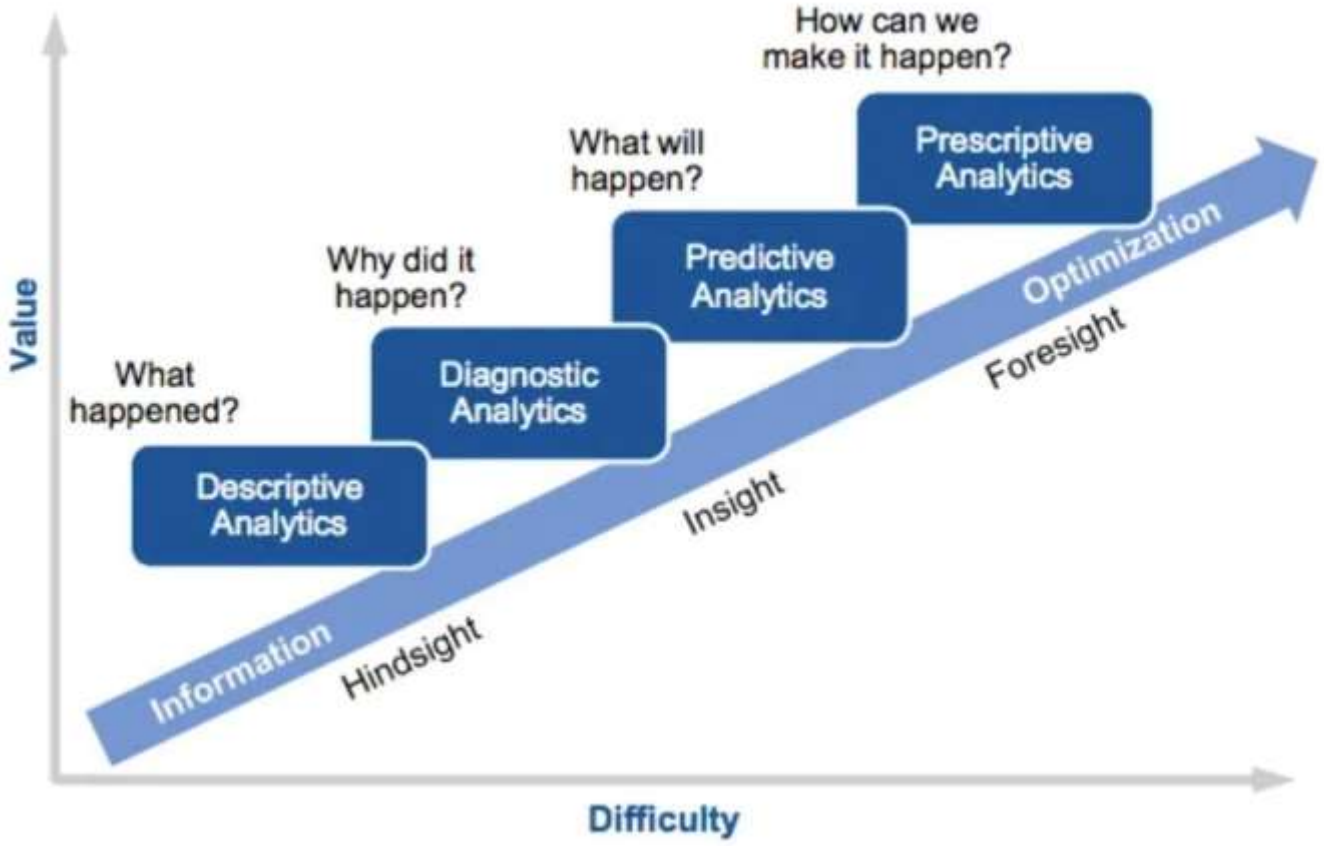
complaints
resolved complaints
CSAT (Cust Satisfaction)

Fraud

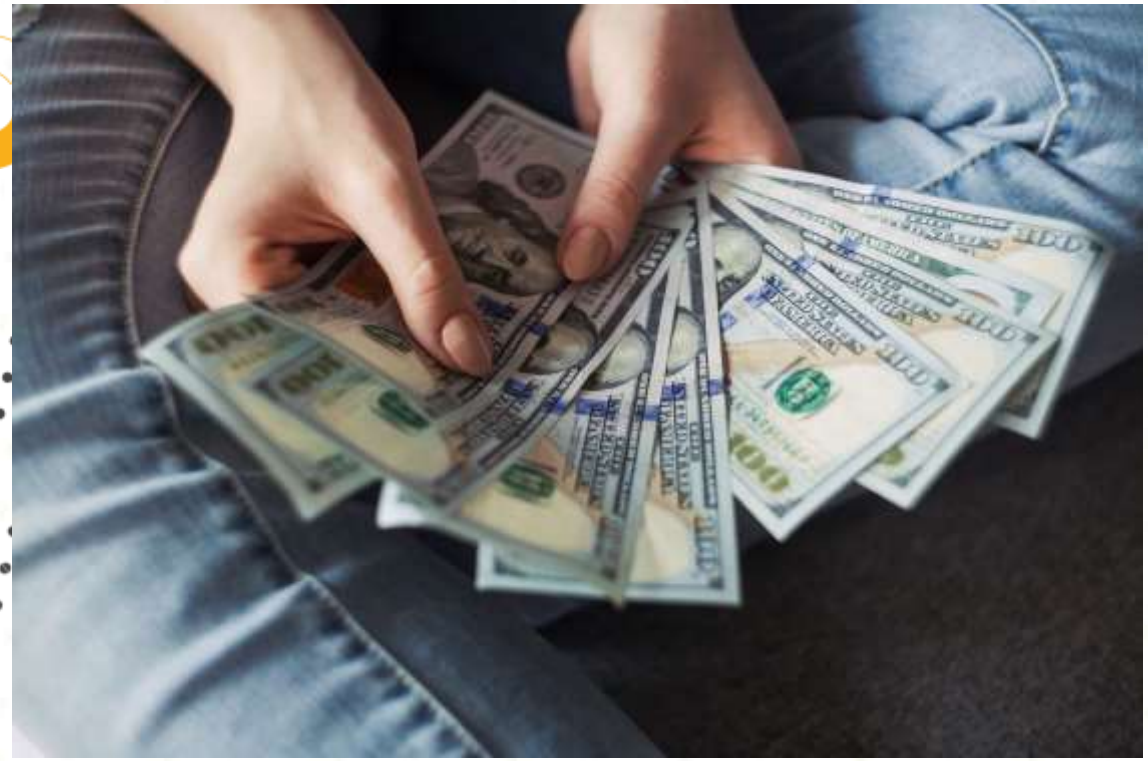
caught frauds
cost loss

[Analytic Approach]

2. Bagaimana kamu bisa menggunakan data untuk menjawab problem tersebut?



[Contoh Kasus: Loan Approval]



Sebuah perusahaan fintech punya **layanan peminjaman modal**. Para customer bisa meminjam sejumlah uang dengan **mengirim pengajuan** secara online.

Selanjutnya **team assessment** akan **memeriksa data customer** dan menentukan apakah customer tersebut **layak** untuk mendapatkan pinjaman.

Seiring berkembangnya perusahaan, customer yang mengajukan pinjaman semakin banyak, dan **kapasitas team assessment sudah tidak cukup** menanganinya.

[Contoh Kasus: Loan Approval]

Business Understanding

Problem : Bagaimana cara membantu team assessment memeriksa pinjaman customer?
Goals : Meningkatkan kecepatan pemeriksaan pengajuan tanpa menambah cost (e.g. menambah agen di team assessment)
Objective : Membuat sistem untuk membantu assessment pinjaman secara otomatis

Business metrics:

- **daily resolved applications** (banyak pengajuan yang berhasil diassess dalam sehari)
- **average resolve time** (rata-rata waktu yang dibutuhkan dari pengajuan masuk hingga assessment selesai)

Analytic Approach

Predictive Analytics: Membuat model yang bisa membantu merekomendasikan apakah pengajuan pinjaman suatu customer diterima/ditolak.

(*Pengajuan ditolak jika ada potensi gagal bayar)

[Data Requirements]

3. Data apa yang kamu butuhkan untuk menjawab problem tersebut?

[Data Collection]

4. Dari mana data tersebut berasal dan bagaimana cara mendapatkannya?

Banyak cara mendapatkan data ...

DigitalSkola
Uncover The World Of Digital Skills With Us

Internal

Database Production
data transaksi
data user
data product

Database Events
data user click
data user page view
data user scroll
tracker infrastruktur

Documents
file excel
kumpulan gambar

Eksternal

Data Public
open data
data repository
public dashboard

Data 3rd Party
data dari vendor
data survey

Scraping /API
website
sosial media

[Contoh Kasus: Loan Approval]

Data Requirements & Data Collection

Data Profil Customer
(dari database)

- Nama
- No KTP
- No HP
- Email
- Gender
- Tanggal lahir
- Alamat
- Status perkawinan
- Jumlah anak
- Pendidikan
- Pekerjaan
- Lama bekerja
- Penghasilan
- ...

Data Pengajuan
(dari database)

- ID Pinjaman (Loan ID)
- Waktu pengajuan
- Jumlah pinjaman
- Jangka waktu
- Tujuan
- Pengajuan Disetujui?
- Status Pinjaman

BI Checking
(external)

- Apakah riwayat pinjaman bagus?

Telekomunikasi
(external)

- No HP prabayar / pascabayar

	LoanID	Gender	Marital Status	Children	Education	Job	Salary	Loan Amount	Loan Term (days)	BI Checking	Decision
0	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
1	N2007	Male	Married	1	S1	Kar. Swasta	9166000	256000	60	not good	REJECT
2	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
3	N2013	Male	Married		D3	Kar. Swasta	5166000	240000	90	good	APPROVE
4	N2017	Male	Single	0	S2	Kar. Swasta	12000000	282000	60		APPROVE
...	...	...	...	...	...	...	...	...	...	...	...
609	N5957	Female	Widowed	0	S1	Kar. Swasta	5800000	142000	60	good	APPROVE
610	N5959	Male	Married	3+	S1	PNS	8212000	80000	30	good	APPROVE
611	N5967				S2	Kar. Swasta	16144000	506000	60		REJECT
612	N5969	Male	Married	2	S1	Kar. Swasta	15166000	374000	60	good	APPROVE
613	N5981	Female	Single	0	S1	Wiraswasta	9166000	266000	60	not good	REJECT

[Data Understanding]
5. Apakah data yang telah kamu kumpulkan sudah representatif?

Lakukan **EDA (Exploratory Data Analysis)!**

Descriptive Statistics

mean, median, mode, min, max, missing values, etc

Correlation Analysis

pairwise correlation, drop highly correlated variables

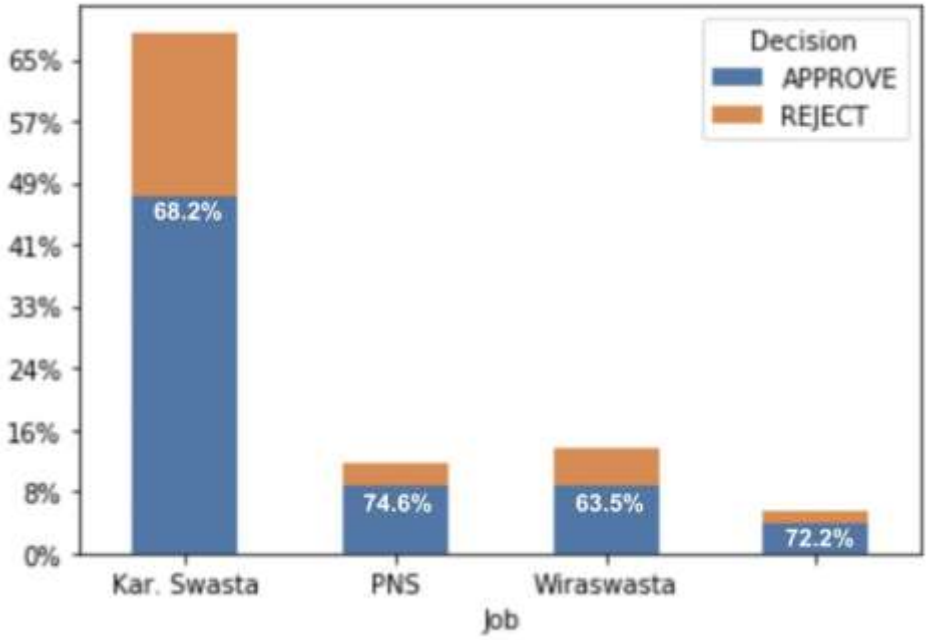
Visualization

lihat pola, trend, & insight

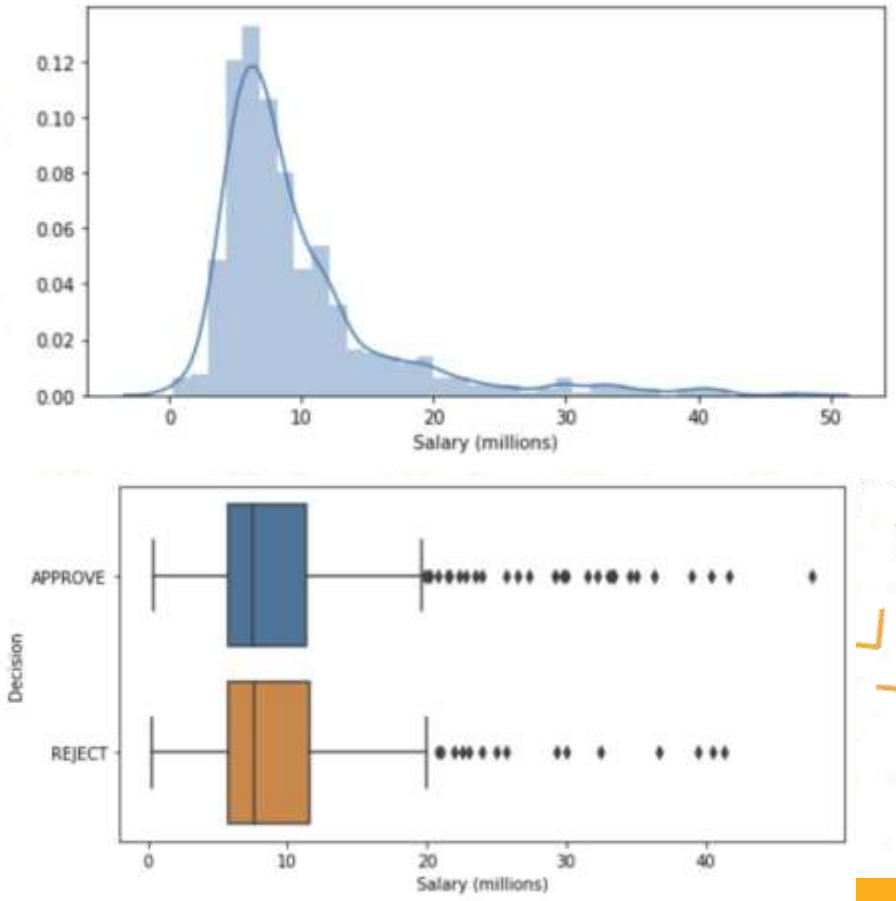
[Contoh Kasus: Loan Approval]

Data Understanding

Atribut: Pekerjaan



Atribut: Penghasilan



- Rata-rata penghasilan peminjam adalah sekitar Rp 9 juta
- Tidak banyak perbedaan distribusi penghasilan antara yang diterima dan ditolak pengajuannya

challenge

- Insight apa lagi yang bisa digali dari dataset loan approval ini?

	LoanID	Gender	Marital Status	Children	Education	Job	Salary	Loan Amount	Loan Term (days)	BI Checking	Decision
0	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
1	N2007	Male	Married	1	S1	Kar. Swasta	9166000	256000	60	not good	REJECT
2	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
3	N2013	Male	Married		D3	Kar. Swasta	5166000	240000	90	good	APPROVE
4	N2017	Male	Single	0	S2	Kar. Swasta	12000000	282000	60		APPROVE
...	...	...	...	...	...	...	...	...	...	...	...
609	N5957	Female	Widowed	0	S1	Kar. Swasta	5800000	142000	60	good	APPROVE
610	N5959	Male	Married	3+	S1	PNS	8212000	80000	30	good	APPROVE
611	N5967				S2	Kar. Swasta	16144000	506000	60		REJECT
612	N5969	Male	Married	2	S1	Kar. Swasta	15166000	374000	60	good	APPROVE
613	N5981	Female	Single	0	S1	Wiraswasta	9166000	266000	60	not good	REJECT

[Data Preparation]

6. Modifikasi apa yang dibutuhkan pada data agar dapat digunakan?

1. Handle Missing Data
2. Handle Invalid Values
3. Remove Duplicates
4. Formatting
5. Feature Engineering

[Contoh Kasus: Loan Approval]

Data Preparation

Membersihkan Data Bermasalah

- Banyak kolom kosong → **hapus**
- Atribut Gender kosong → **diisi 'Male'**
- Atribut Gender 'Perempuan' → **diisi 'Female'**
- Atribut Children kosong → **diisi 0**
- Atribut Salary kosong → **diisi rata-rata/estimasi**
- Ada data duplikat → **hapus, sisakan 1**
- …

Formatting & Feature Engineering

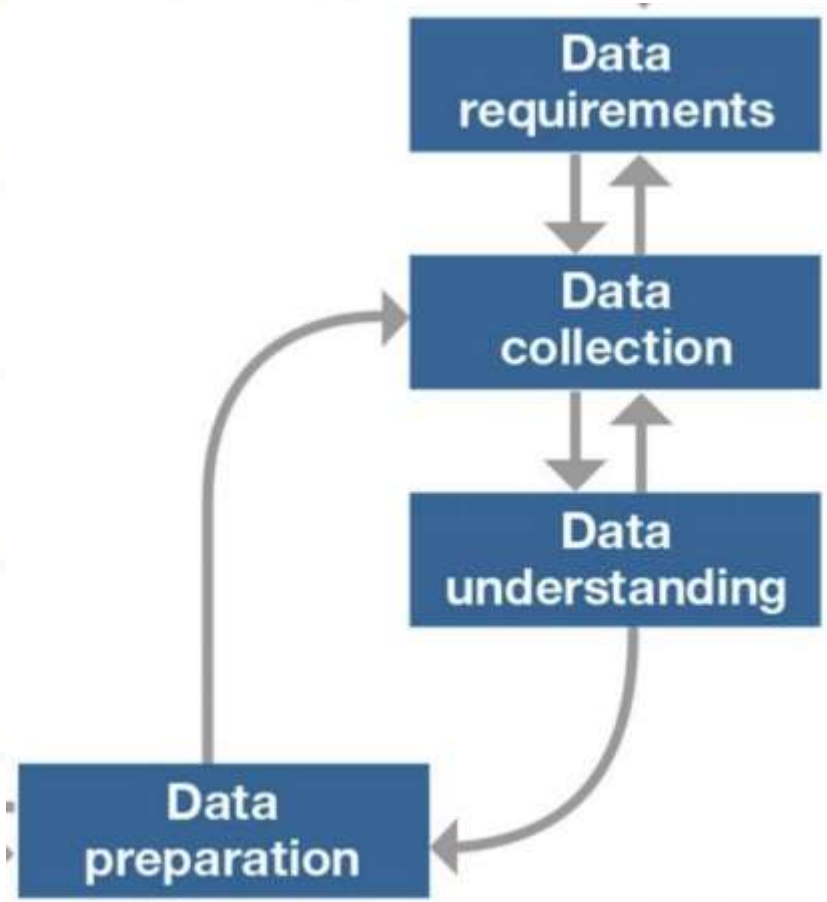
- **BI Checking** → nilai dijadikan 1 atau 0
- **Loan Term** → dari hari menjadi bulan

- **Age** → hitung dari tanggal lahir
- **# previous application** → hitung dari data historical
- **Installment** → hitung dari Pinjaman & Jangka Waktu
- **HP Number Type** (prabayar/pascabayar)

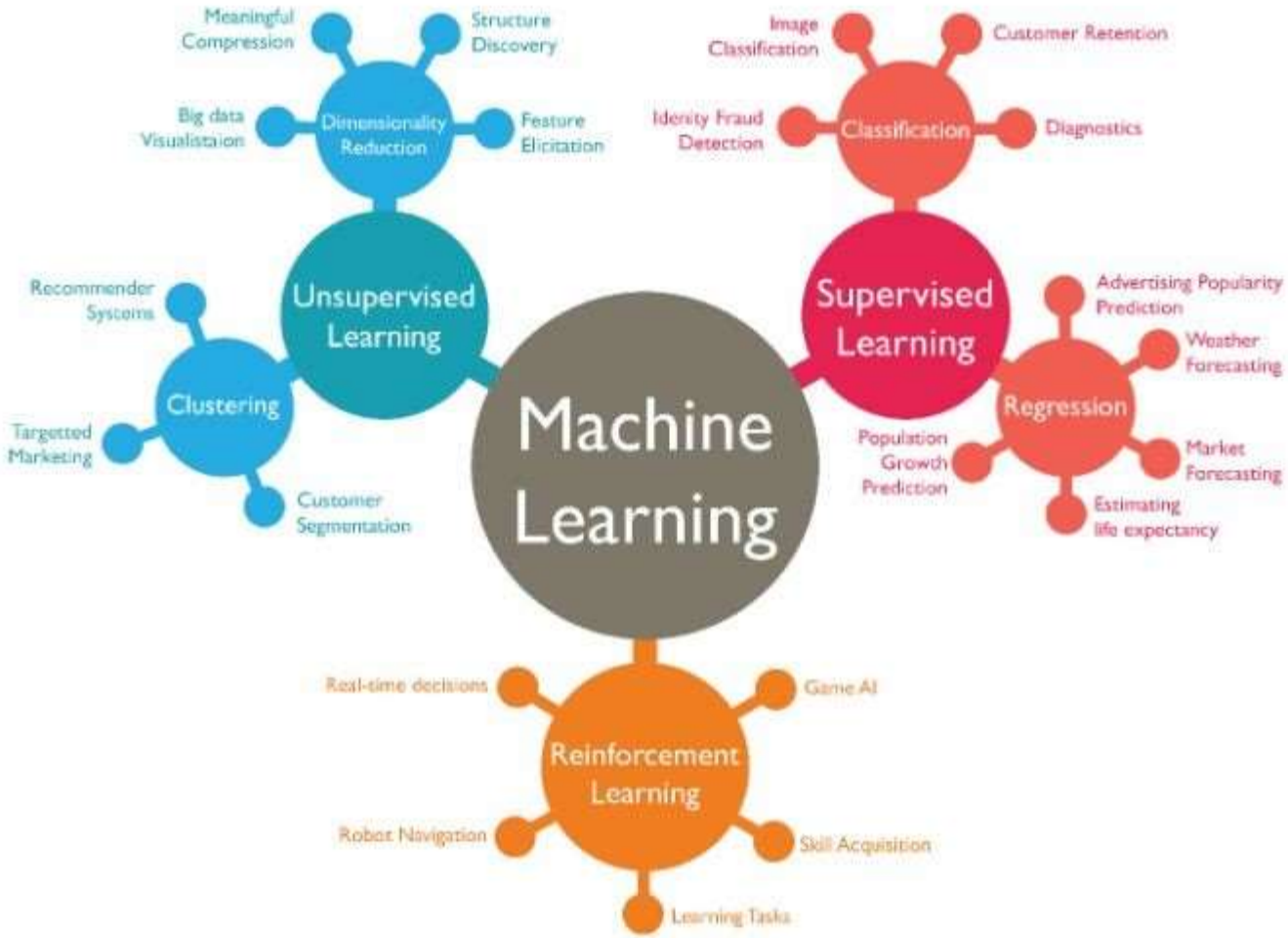
	LoanID	Gender	Marital Status	Children	Education	Job	Salary	Loan Amount	Loan Term (days)	BI Checking	Decision
0	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
1	N2007	Male	Married	1	S1	Kar. Swasta	9166000	256000	60	not good	REJECT
2	N2005	Male	Single	0	S1	PNS	11698000	100000	60	good	APPROVE
3	N2013	Male	Married		D3	Kar. Swasta	5166000	240000	90	good	APPROVE
4	N2017	Male	Single	0	S2	Kar. Swasta	12000000	282000	60		APPROVE
...	...	...	...	...	...	...	...	...	...	...	...
609	N5957	Female	Widowed	0	S1	Kar. Swasta	5800000	142000	60	good	APPROVE
610	N5959	Male	Married	3+	S1	PNS	8212000	80000	30	good	APPROVE
611	N5967				S2	Kar. Swasta	16144000	506000	60		REJECT
612	N5969	Male	Married	2	S1	Kar. Swasta	15166000	374000	60	good	APPROVE
613	N5981	Female	Single	0	S1	Wiraswasta	9166000	266000	60	not good	REJECT

Bekerja Dengan Data!

Lakukan hingga data siap digunakan untuk modeling



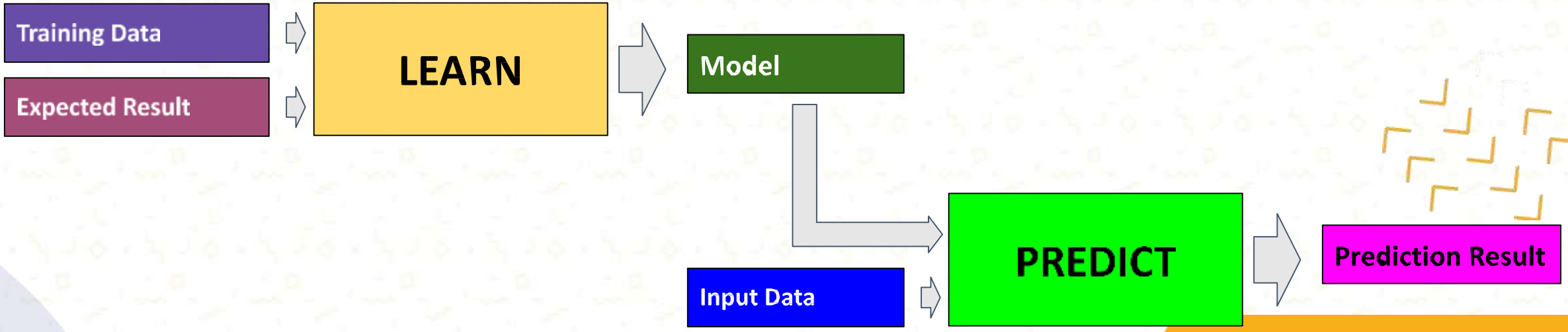
[Modeling]
7. Bagaimana memodelkan data agar dapat menyelesaikan problem?



Conventional vs Machine Learning

Conventional Modelling

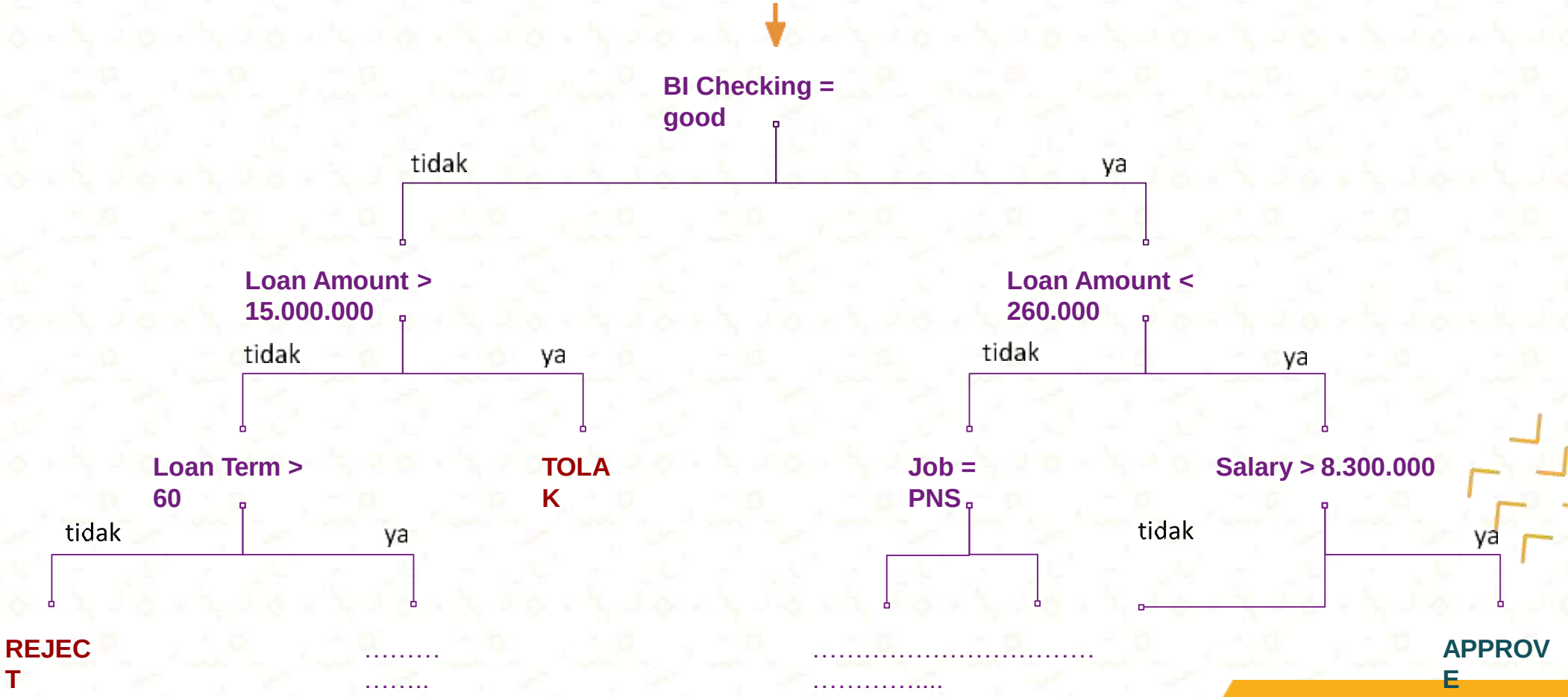
Machine Learning Modelling



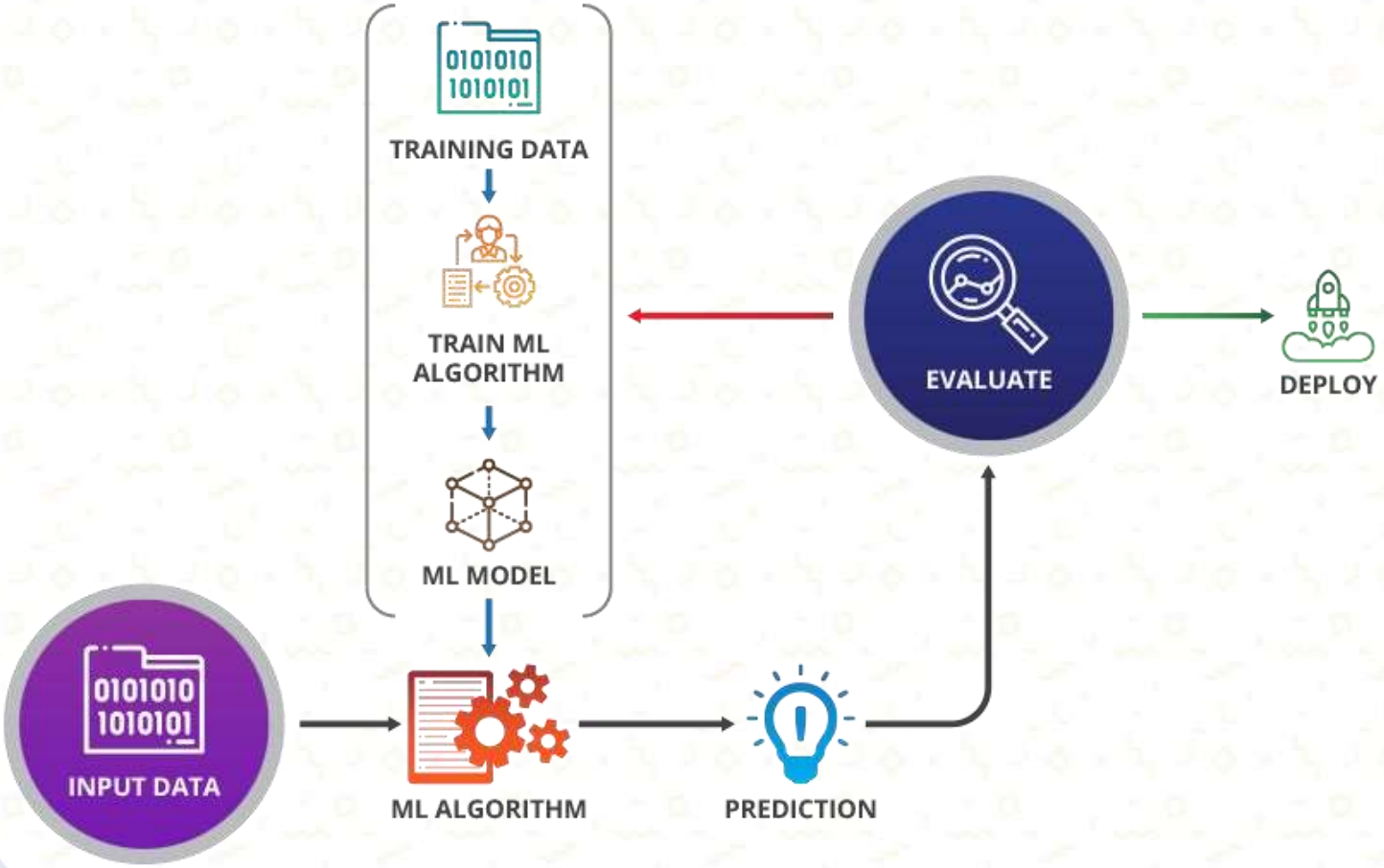
[Contoh Kasus: Loan Approval]

Modelling

Contoh Model Menggunakan Algoritma Decision Tree



[Evaluation]
8.Apakah model sudah cukup baik untuk menyelesaikan problem?

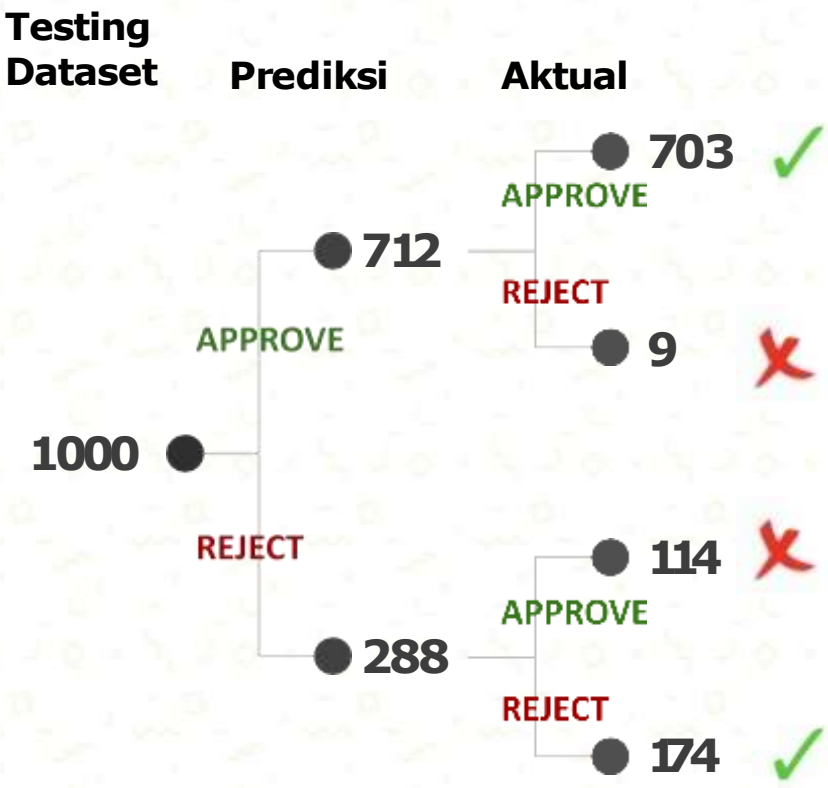


Evaluation Metrics (tergantung problem):

- Accuracy
- Precision
- RMSE
- Significance
- …

[Contoh Kasus: Loan Approval]

Evaluation



Confusion Matrix

Aktual \ Prediksi	APPROVE	REJECT
APPROVE	703	114
REJECT	9	174

Accuracy = (703+174)/1000 = **87.7%**

Precision = 703/(703+9) = **98.7%**

[Contoh Kasus: Loan Approval]

A

Actual \ Prediction	APPROVE	REJECT
APPROVE	703	114
REJECT	9	174

B

Actual \ Prediction	APPROVE	REJECT
APPROVE	714	103
REJECT	15	168

C

Actual \ Prediction	APPROVE	REJECT
APPROVE	655	162
REJECT	10	173

D

Actual \ Prediction	APPROVE	REJECT
APPROVE	788	29
REJECT	127	56

Model	Accuracy	Precision
A	87.7%	98.7%
B	88.2%	97.9%
C	82.8%	98.5%
D	84.4%	86.1%

[Deployment]
9. Dapatkah kita implementasi modelnya?

[Contoh Kasus: Loan Approval]

Deployment

Kemungkinan Skenario:

- **Full automation**
  Pengajuan langsung diterima / ditolak berdasarkan output dari model

- **Auto-reject**
  Pengajuan yang kemungkinan jelek langsung ditolak.
  Jika tidak, perlu dicek manual dulu oleh team assessment

- **Partial Auto-reject & Auto-approve**
  Pengajuan yang kemungkinan jelek langsung ditolak.
  Pengajuan yang kemungkinan tinggi bagus langsung diterima.
  Jika masih 'abu-abu', baru dicek manual oleh team assessment

[Feedback]
10. Apakah ada feedback yang konstruktif?

Feedback bisa dari mana saja…

- Dari monitoring dashboard (performance dengan data real-time)
- Dari end user
- Dari business stakeholders
- Dari engineering
- etc.

[Contoh Kasus: Loan Approval]

Feedback

Metrics Impact

Business Metrics	Before	After
daily resolved applications	10.000	30.000
average resolved time	50 hours	1 hour

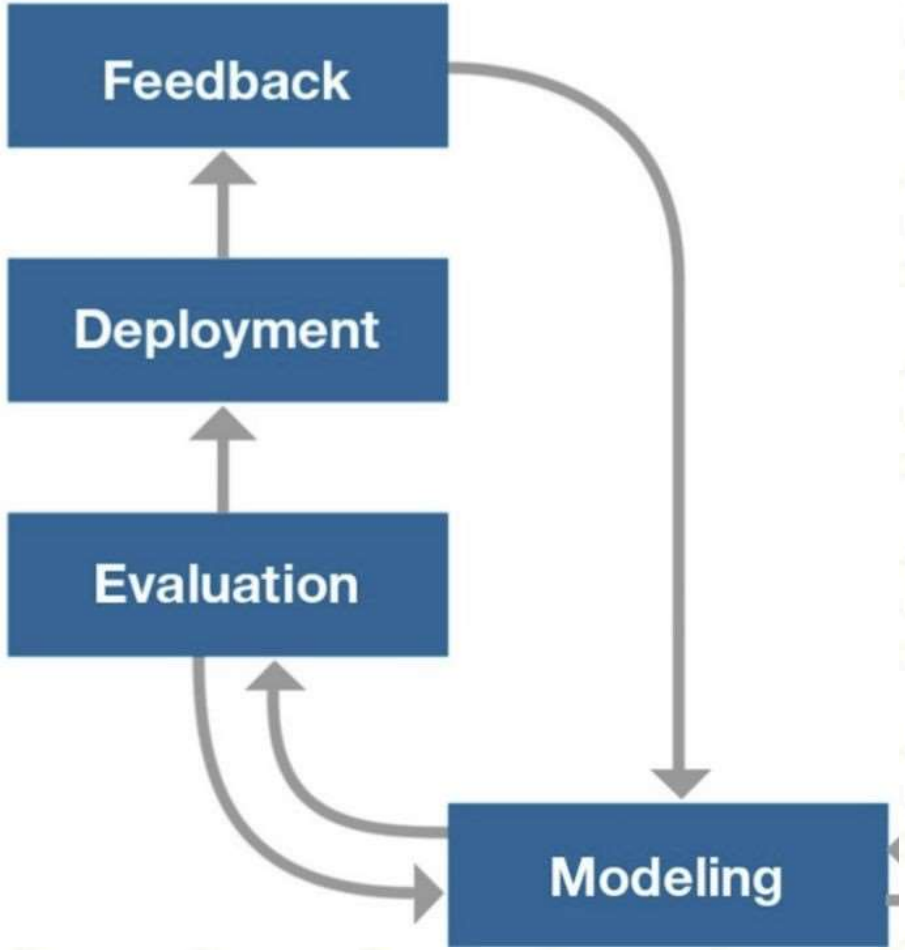
Model Performance in Production

Accuracy : 81.4%
Precision : 92.7%

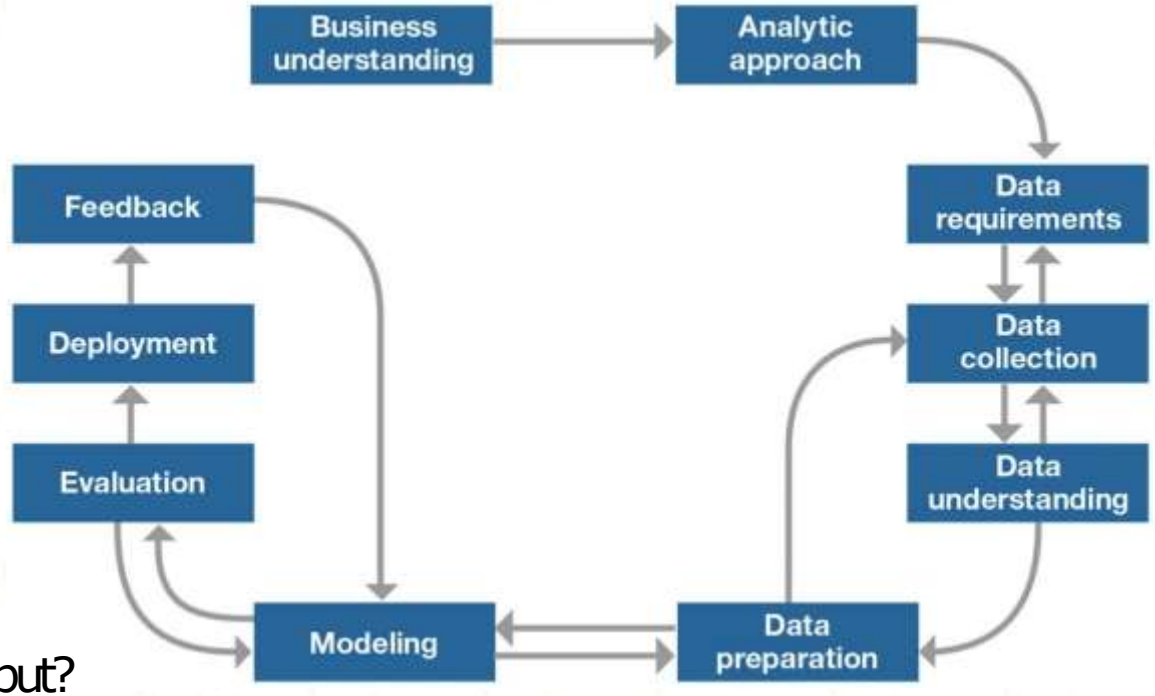
Kalau dirasa terlalu drop, perlu improve model-nya lagi

Membangun solusi!

Lakukan hingga model yang dibuat memberikan hasil yang sesuai harapan



Recap!



Perencanaan

1. Problem apa yang ingin kamu selesaikan?
2. Bagaimana kamu bisa menggunakan data untuk menjawab problem tersebut?

Bekerja dengan data

3. Data apa yang kamu butuhkan untuk menjawab problem tersebut?
4. Dari mana data tersebut berasal dan bagaimana cara mendapatkannya?
5. Apakah data yang telah kamu kumpulkan sudah representatif?
6. Modifikasi apa yang dibutuhkan pada data agar dapat digunakan?

Membangun solusi

7. Bagaimana memodelkan data agar dapat menyelesaikan problem?
8. Apakah model sudah cukup baik untuk menyelesaikan problem?
9. Dapatkah kita implementasi modelnya?
10. Apakah ada feedback yang konstruktif?

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU