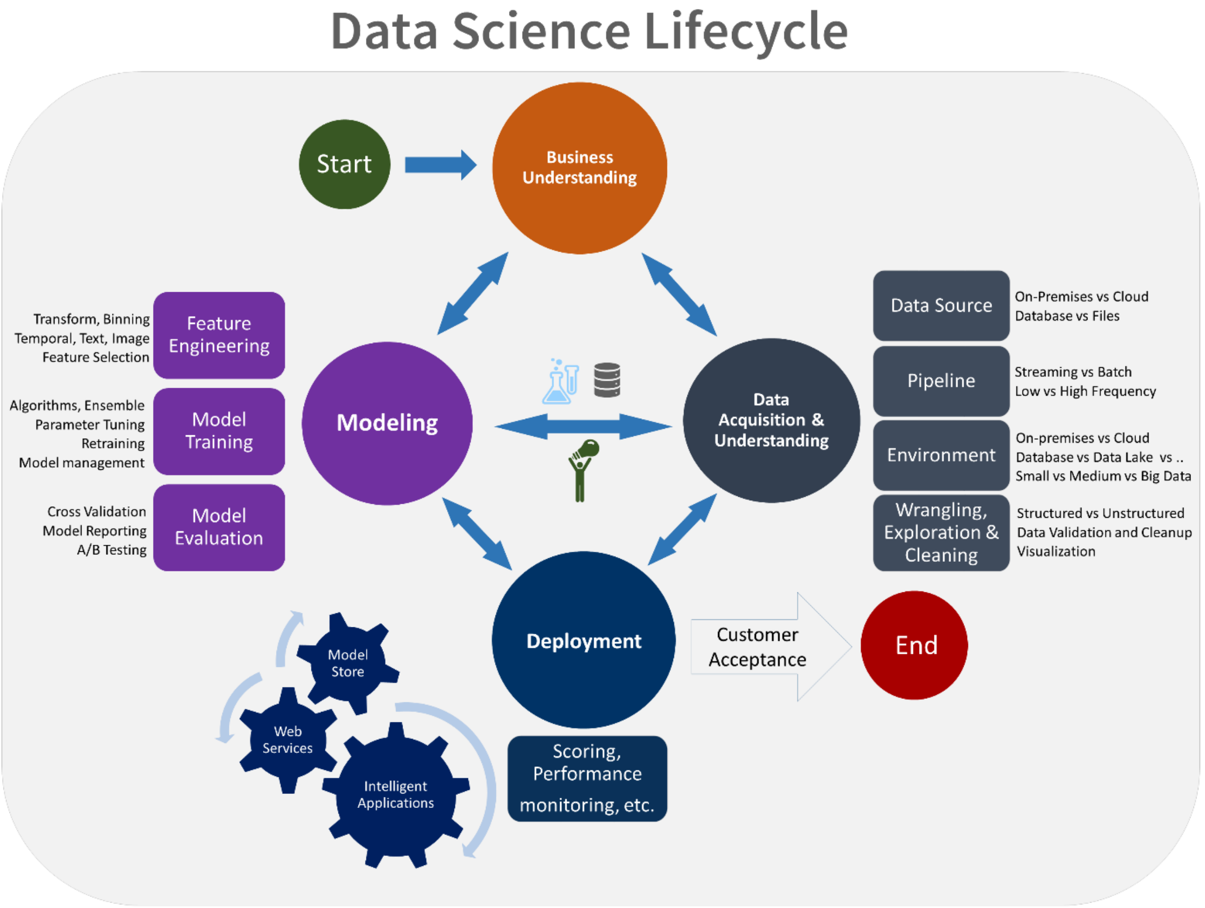
Data Science Project

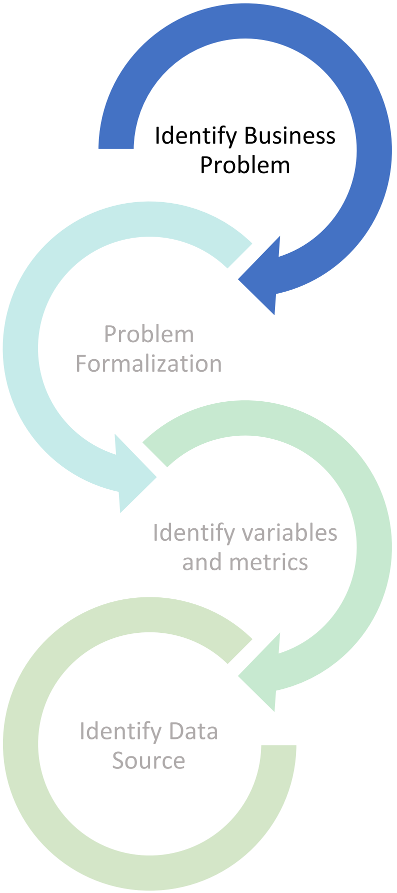
The life cycle

Big Picture



Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Business Understanding



Example

Goal:

To identify

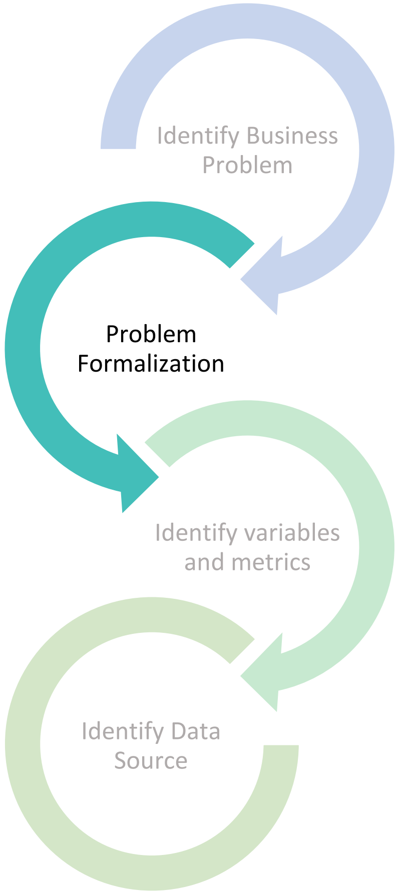
- *What data to be used*
- *Where are data*
- *How to measure*

Scenario: House Mortgages

House price assessment takes approximately 7-14 days which takes time and cost resulting bad customer experience

Problem: *How to estimate the price faster with acceptable accuracy?*

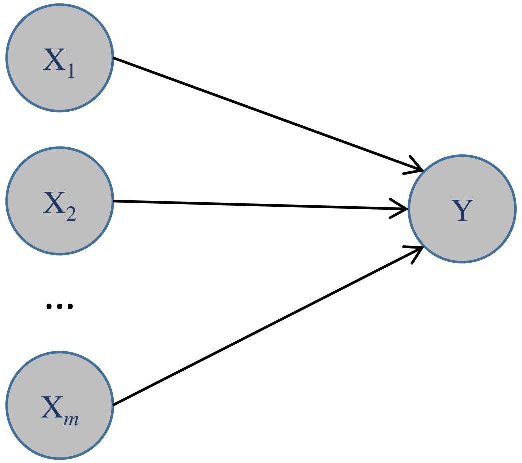
Business Understanding



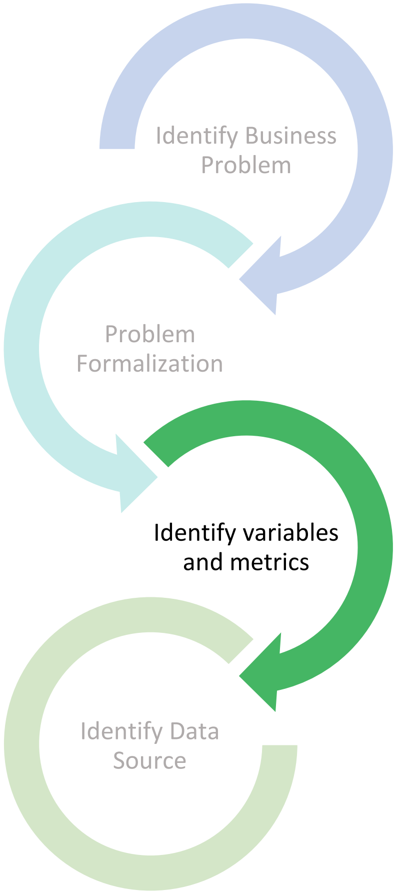
Example

Problem settings:
Given a house profile, estimate its price

Formalization: *Regression Problem*



Business Understanding



Example

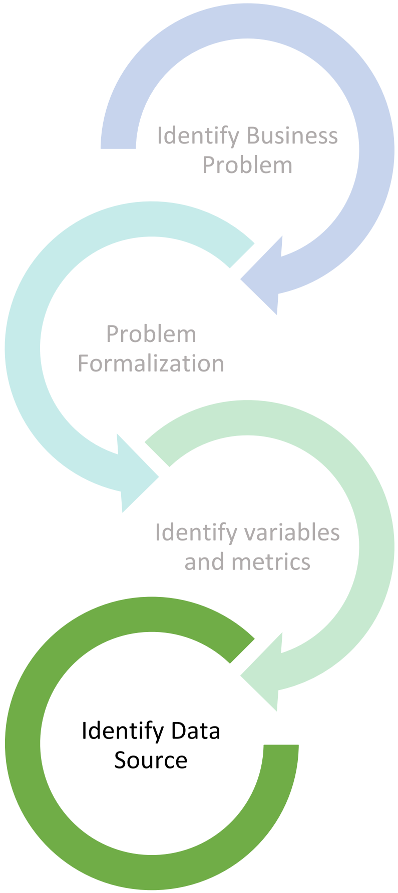
Variables

- no. of room
- no. of floor
- year (how old)
- Area (size)
- Location (land price per acre)
- Distance from important landmark, e.g. school, hospital, … etc

Metrics

- Model accuracy
- Turnaround time
- Cost reduction

Business Understanding



Example

Data sources

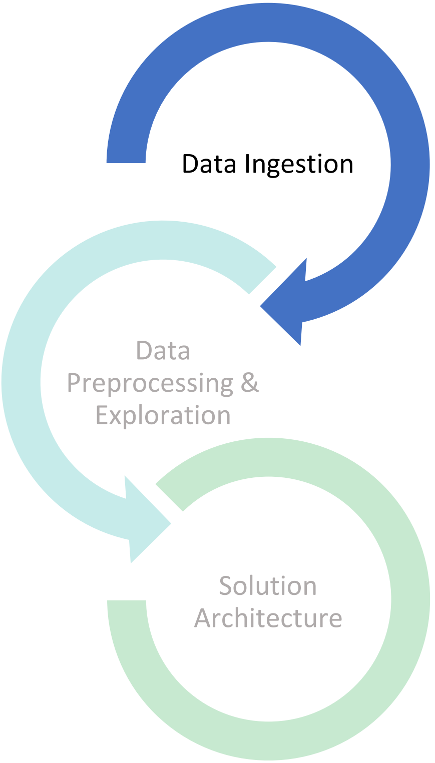
- Internal Data, e.g. Loan database
- External Data, e.g. Geo database, DOLdb

Artifacts

- Charter document[1]
- Data source
- Data Dictionary

[1] Example of Charter Document: https://github.com/Azure/Azure-TDSP-ProjectTemplate/blob/master/Docs/Project/Charter.md

Data Acquisition and Understanding

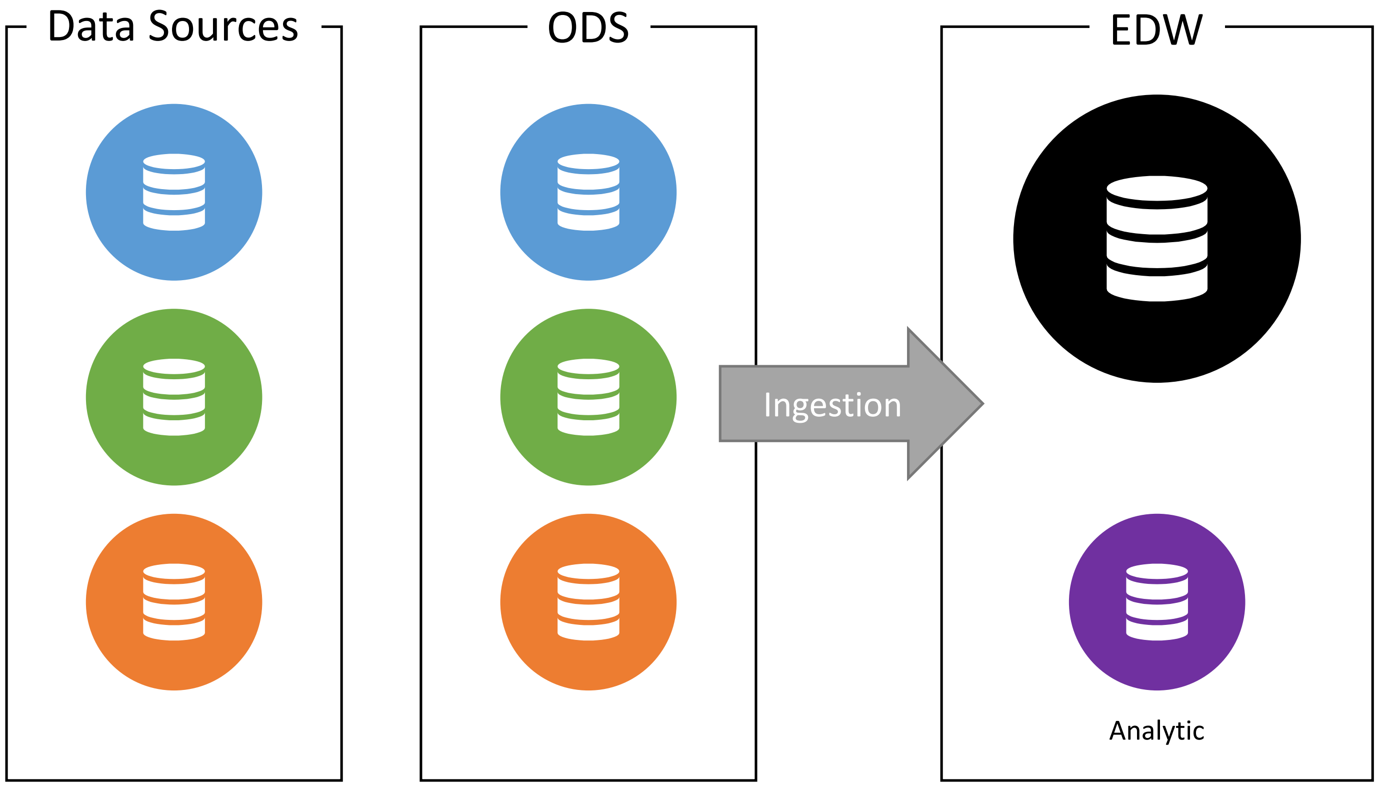


Goal

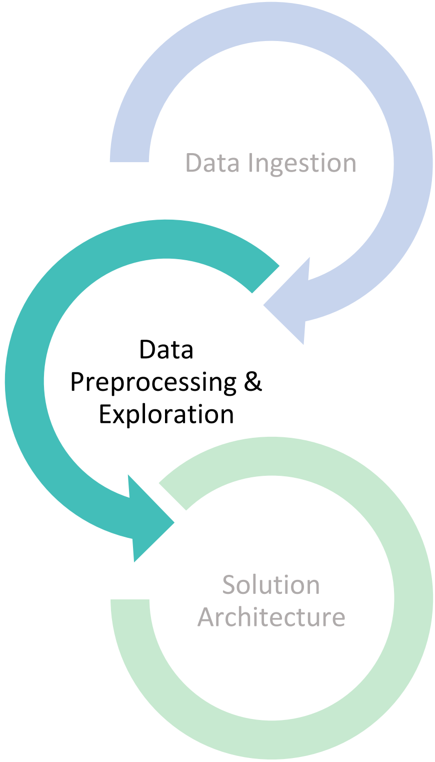
- *To produce high quality data*
- *To ingest data from operation to analytic environment*
- *To develop solution architecture*

Data Acquisition and Understanding

- Data Ingestion



Data Acquisition and Understanding

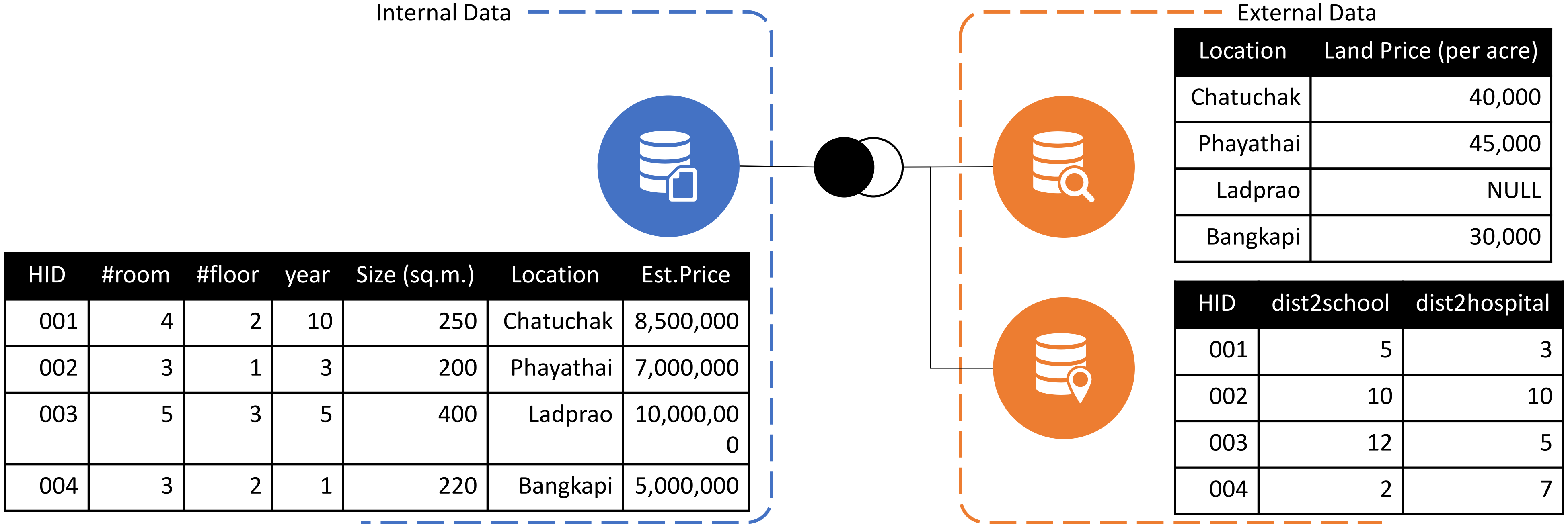


Major Data Preprocessing Tasks

- *Data cleansing, e.g. missing value handling*
- *Data transformation, e.g. rescaling, normalization*
- *Data reduction, e.g. data sampling*
- *Data discretization, e.g. continuous to category conversion*
- *Text cleansing, e.g. inconsistent delimiters*

Data Acquisition and Understanding

- Data Preprocessing and Data Exploration
  - From house mortgage scenario



HID	#room	#floor	year	Size (sq.m.)	Location	Est.Price
001	4	2	10	250	Chatuchak	8,500,000
002	3	1	3	200	Phayathai	7,000,000
003	5	3	5	400	Ladprao	10,000,000
004	3	2	1	220	Bangkapi	5,000,000

Location	Land Price (per acre)
Chatuchak	40,000
Phayathai	45,000
Ladprao	NULL
Bangkapi	30,000

HID	dist2school	dist2hospital
001	5	3
002	10	10
003	12	5
004	2	7

Data Acquisition and Understanding

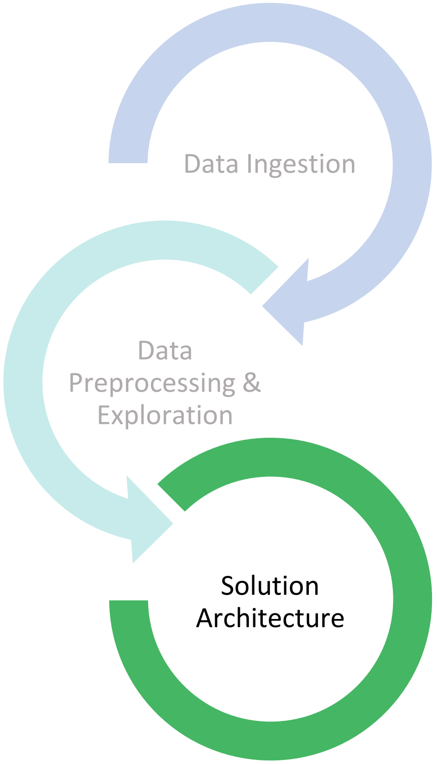
- Data Preprocessing and Data Exploration

#room	#floor	year	Size (sq.m.)	Land price	Dist2school (km)	Dist2hospital (km)	Est.Price
4	2	10	250	40,000	5	3	8,500,000
3	1	3	200	45,000	10	10	7,000,000
5	3	5	400	NULL	12	5	10,000,000
3	2	1	220	30,000	2	7	5,000,000

🟦 Internal Data

🟩 External Data

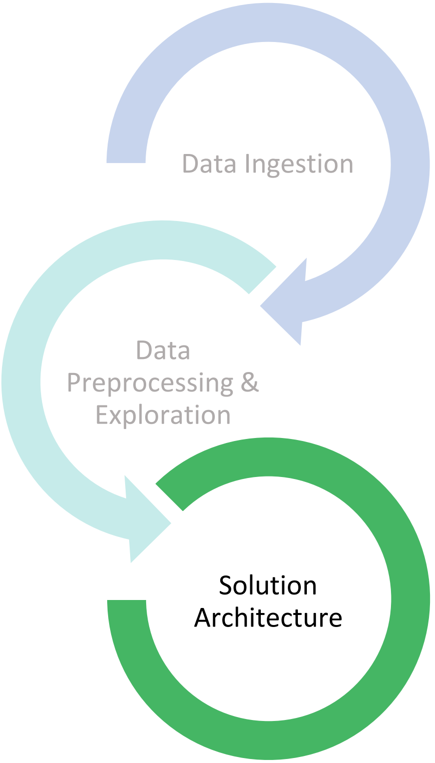
Data Acquisition and Understanding



Batch:
Standard architecture for data warehouse such as *Data mart*

Stream:
Data flows continuously from the data sources. This idea is called *Data Lake*.

Data Acquisition and Understanding



Data Ingestion

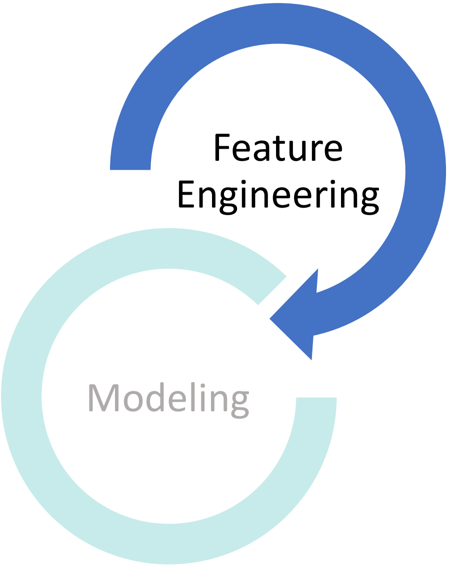
Data Preprocessing & Exploration

Solution Architecture

Data Mart:
Just like a store of bottled water – cleansed and packaged and structured for easy consumption

Data Lake:
Similar to a large body of water in a more natural state. Various users of the lake can come to examine, dive in, or take samples

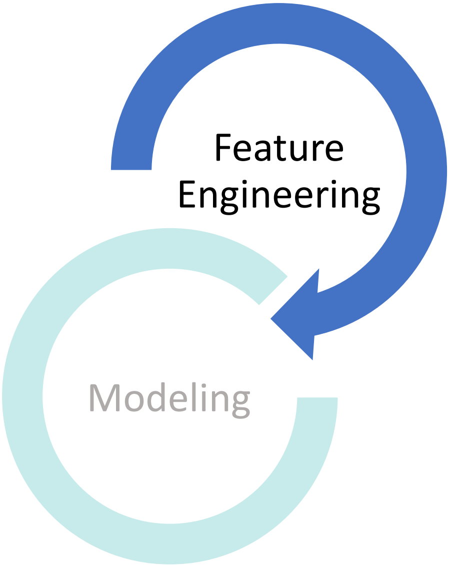
Modeling



Goal

- *To create a list of feature vectors from raw data*
- *To create a machine learning model*

Modeling



Structured Data

- Pick all relevant variables to the target class
- Data Preprocessing

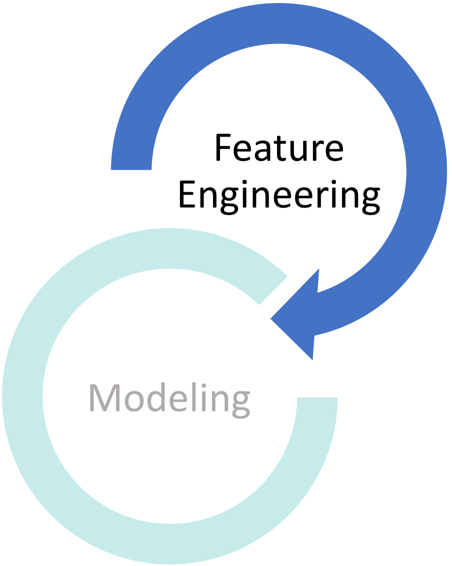
Weight	Height	Location	Diabetes
50	155	Bangkok	No
60	165	Bangkok	Yes
70	160	Nakon Ratchasima	No
80	150	Nakon Ratchasima	Yes
90	168	Nakon Ratchasima	Yes

✓ ✓ ✗

Modeling

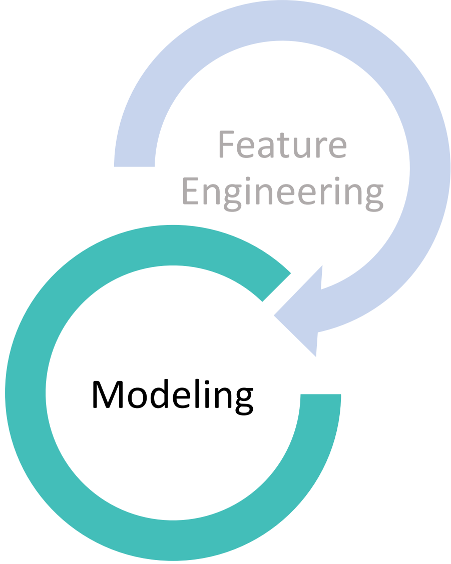
Structured Data

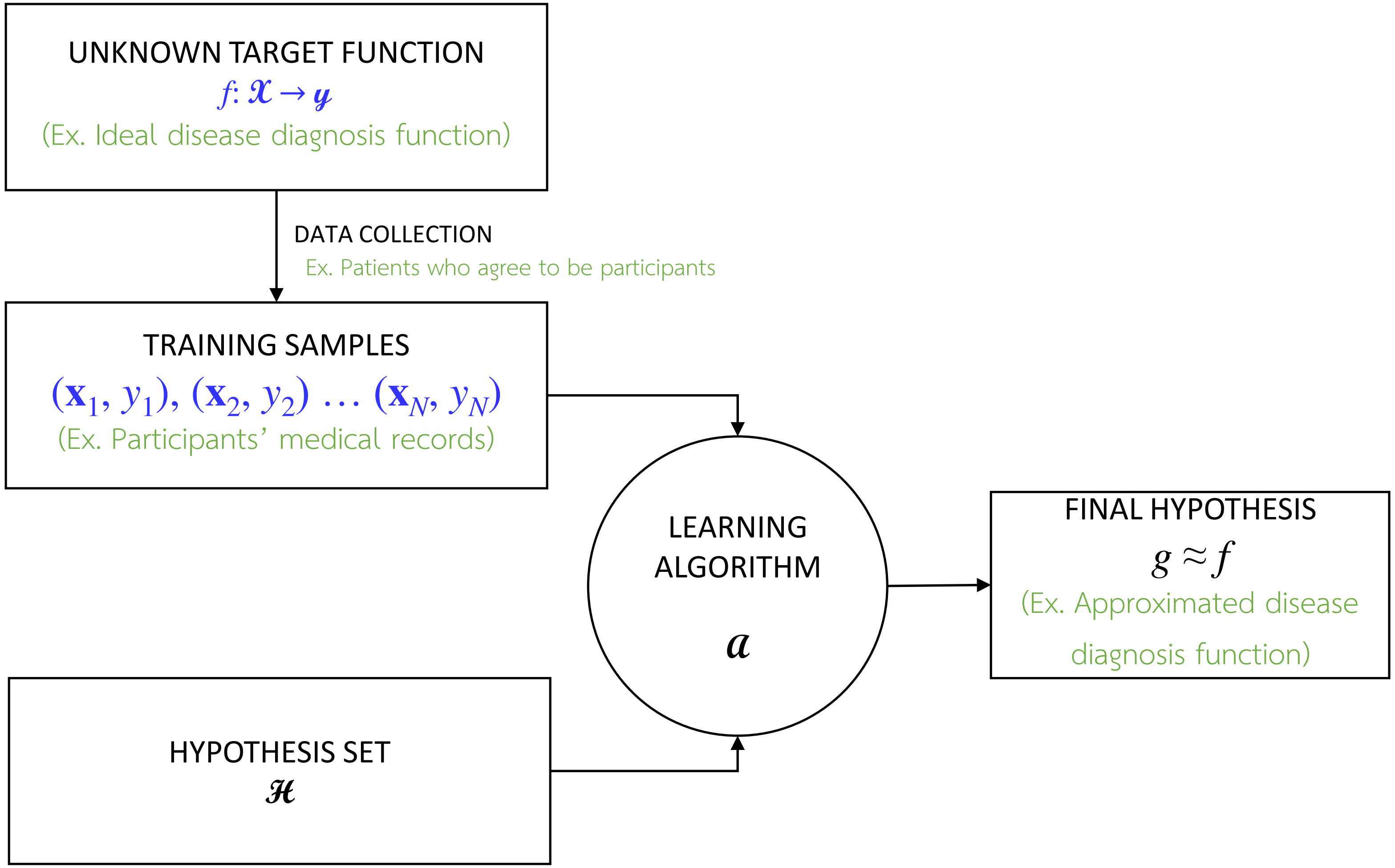
- Pick all relevant variables to the target class
- Data Preprocessing



Weight	Height	BMI	Diabetes
50	155	$50/1.55^2 = 20.8$	No
60	165	$60/1.65^2 = 22$	Yes
70	160	$70/1.60^2 = 27.3$	No
80	150	$80/1.50^2 = 35.6$	Yes
90	168	$90/1.68^2 = 31.9$	Yes

Modeling





UNKNOWN TARGET FUNCTION

$f: \mathcal{X} \rightarrow \mathcal{Y}$

(Ex. Ideal disease diagnosis function)

DATA COLLECTION

Ex. Patients who agree to be participants

TRAINING SAMPLES

$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_N, y_N)$

(Ex. Participants' medical records)

LEARNING ALGORITHM

$\mathcal{A}$

FINAL HYPOTHESIS

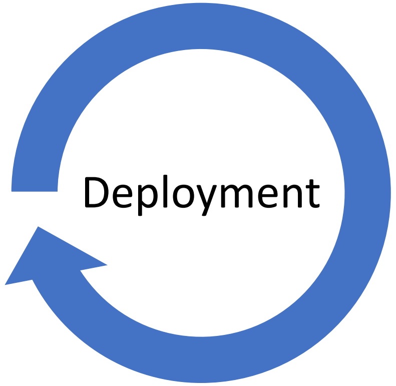
$g \approx f$

(Ex. Approximated disease diagnosis function)

HYPOTHESIS SET

$\mathcal{H}$

Deployment

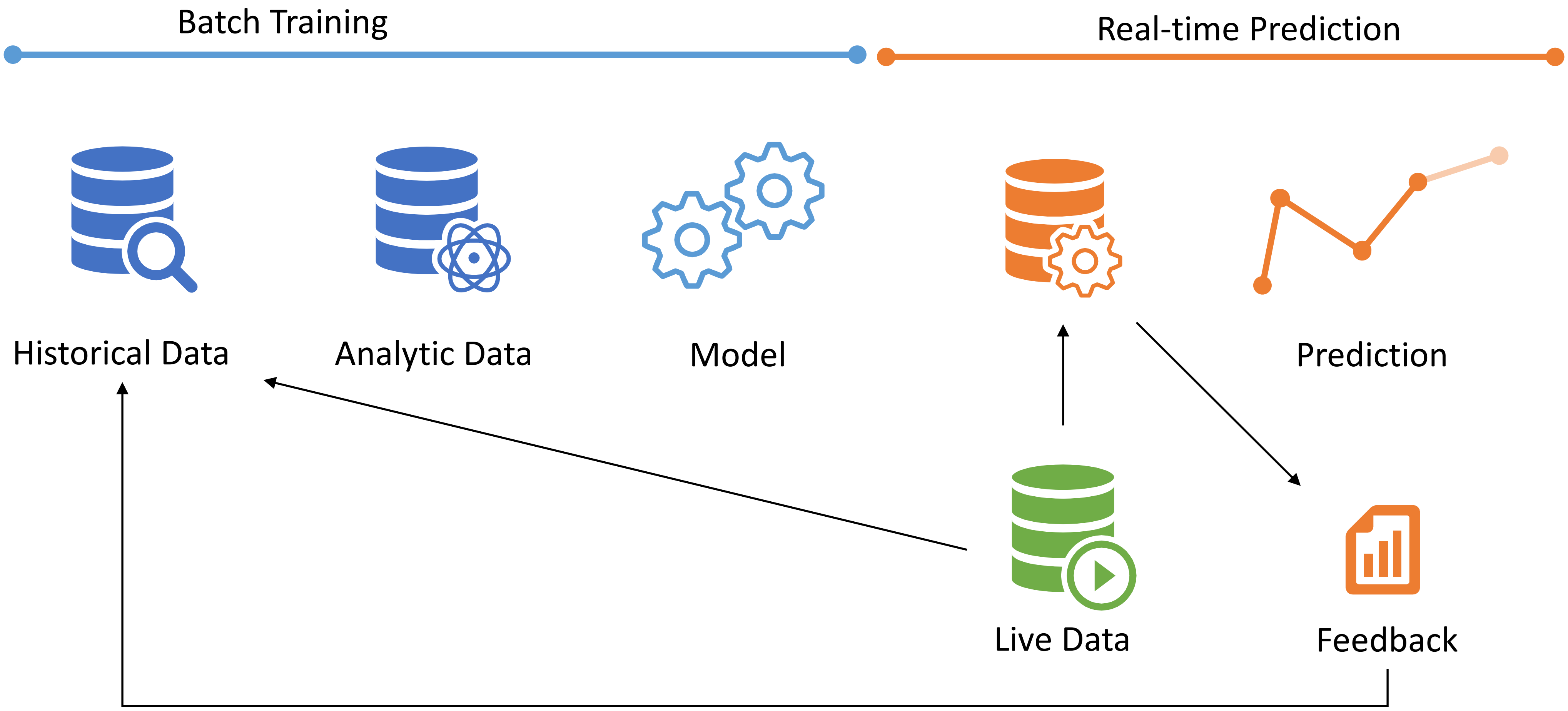


Deployment

Goal

- *Deploy models with a data pipeline to a production or production-like environment for final user acceptance*
- *Tracking model performance and improving if required*

Deployment


Batch Training
Real-time Prediction
Historical Data
Analytic Data
Model
Prediction
Live Data
Feedback