STelligence

Data Preparation for AI

Veerasit Kaewbundit (Leo)

AI Engineer

STELLIGENCE Co., Ltd.

linkedin.com/in/veerasit-kaewbundit

veerasit.k@stelligence.com

Veerasit Kaewbundit

AI Engineer

veerasit.k@stelligence.com

linkedin.com/in/veerasit-kaewbundit

Education

- **Bachelor's Degree, Industrial Engineering with First Class Honours**
  Kasetsart University (2023)

Work Experience

- **AI Engineer**
  STELLIGENCE Co., Ltd. (2023 – Present)
- **Data Scientist & Data Engineer**
  STELLIGENCE Co., Ltd. (2023)
- **Researcher (Data Science and Data Analytics)**
  Department of Industrial Engineering, KU (2022 – Present)
- **Teaching Assistant (Applied Mathematics for Engineers, Industrial Safety)**
  Department of Industrial Engineering, KU (2021 – 2023)
- **Data Analyst (Internship) & Data Engineering Consultant (Part-time)**
  STELLIGENCE Co., Ltd. (2022)

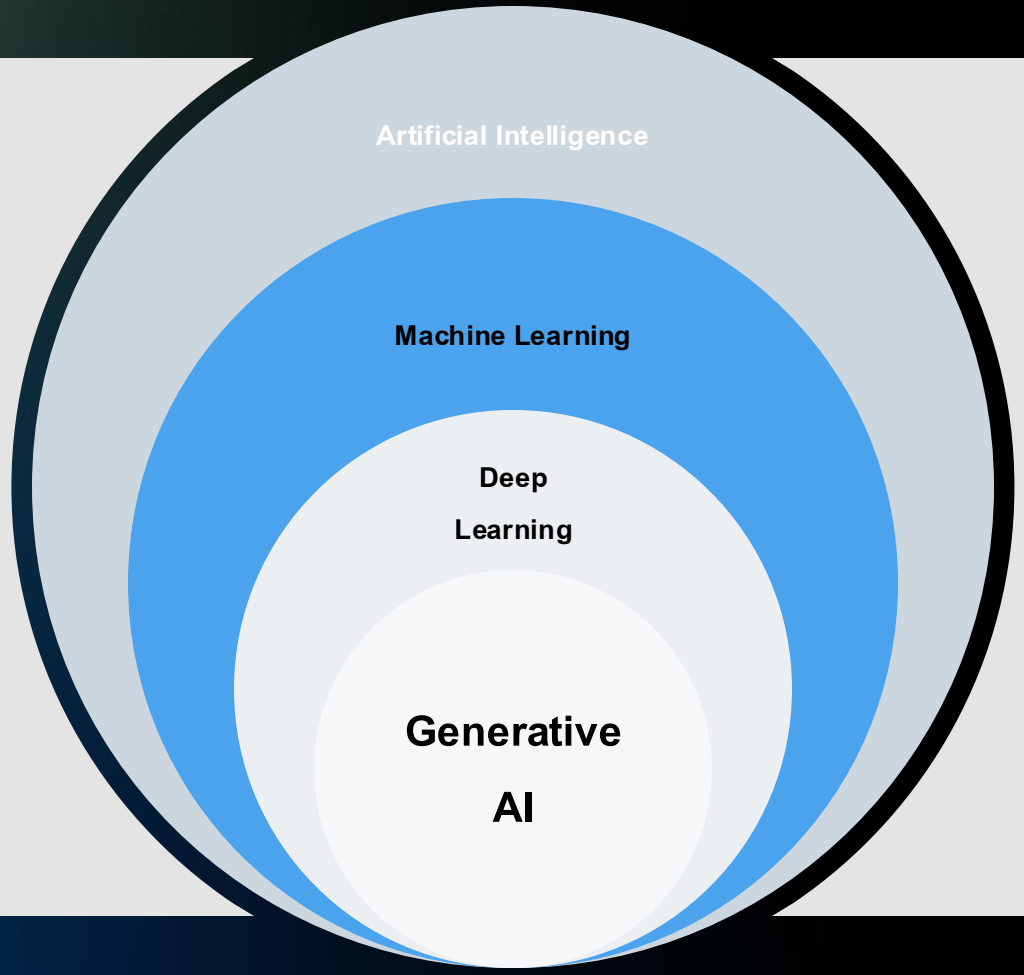
Certification

- Introduction to DevOps – IBM (2025)
- Graph Data Science Certification – Neo4j (2024)
- SUPER AI ENGINEER SEASON 4 Foundation AI Theory – AIAT (2024)
- Certified Professional – Neo4j (2023)
- Generative AI with Large Language Models – DeepLearning.AI & AWS (2023)

Publication

- A Spatiotemporal Aerosol Optical Depth Forecasting in Thailand using Deep Learning. 2025 14th International Conference on Computing and Pattern Recognition (ICCPR 2025), To be in 2025.
- A spatiotemporal deep learning ensemble for multi-step PM2.5 prediction: A case study of Bangkok metropolitan region in Thailand. Atmospheric Pollution Research (Q1), 16(3), 102406.

The journey continues with generative AI



1956

Artificial Intelligence

The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997

Machine Learning

Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012

Deep Learning

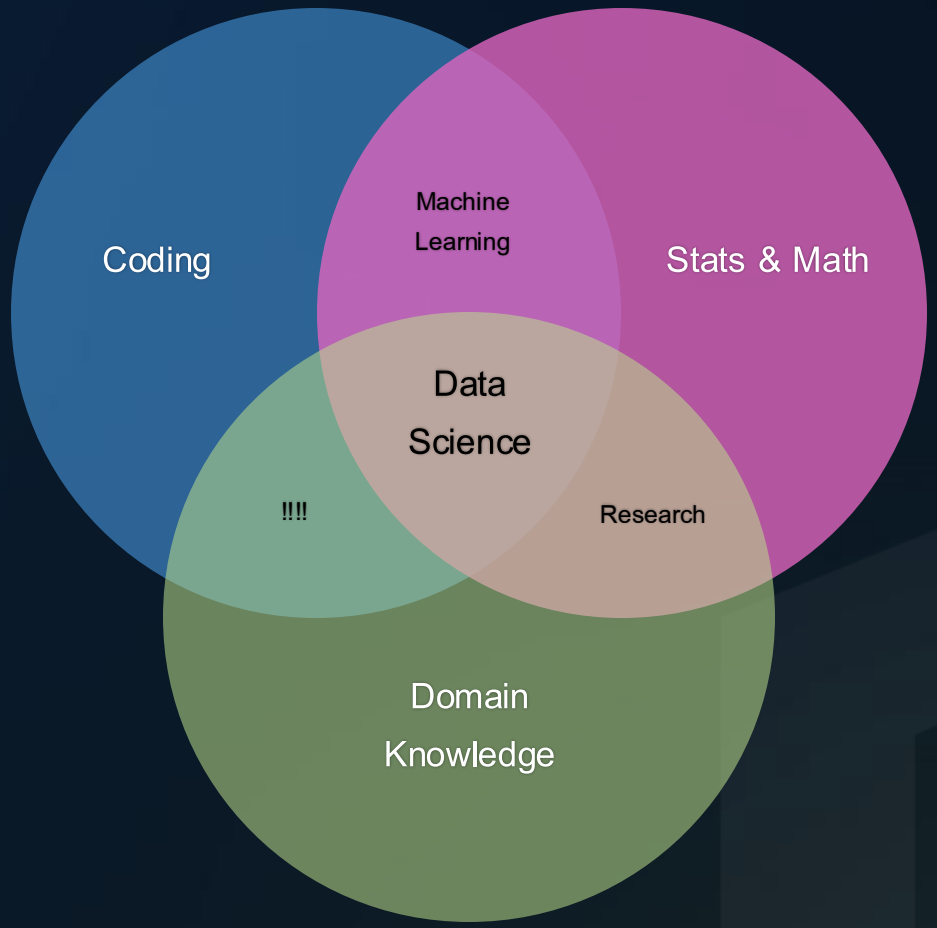
A machine learning technique in which layers of neural networks are used to process data and make decisions.

2021

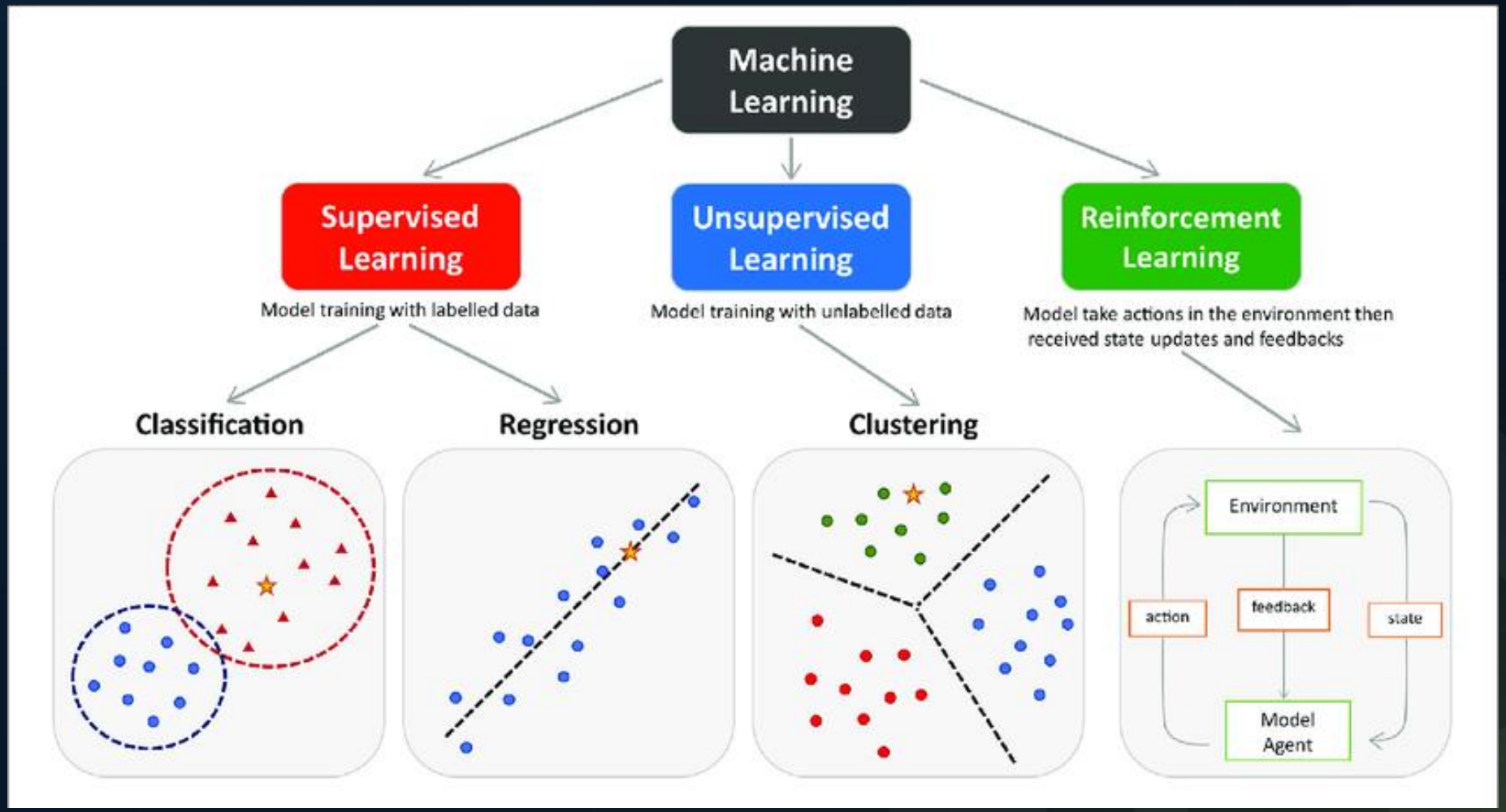
Generative AI

Create new written, visual, and auditory content given prompts or existing data.

Components of Data Science



Type of Machine Learning



Terminology

Supervised Vs Unsupervised Learning, Explained

Features
Predictors
Independent Variables

Samples
Data Points

Supervised

X_1	X_2	X_p	Y

Target

Label
Response
Dependent Variable

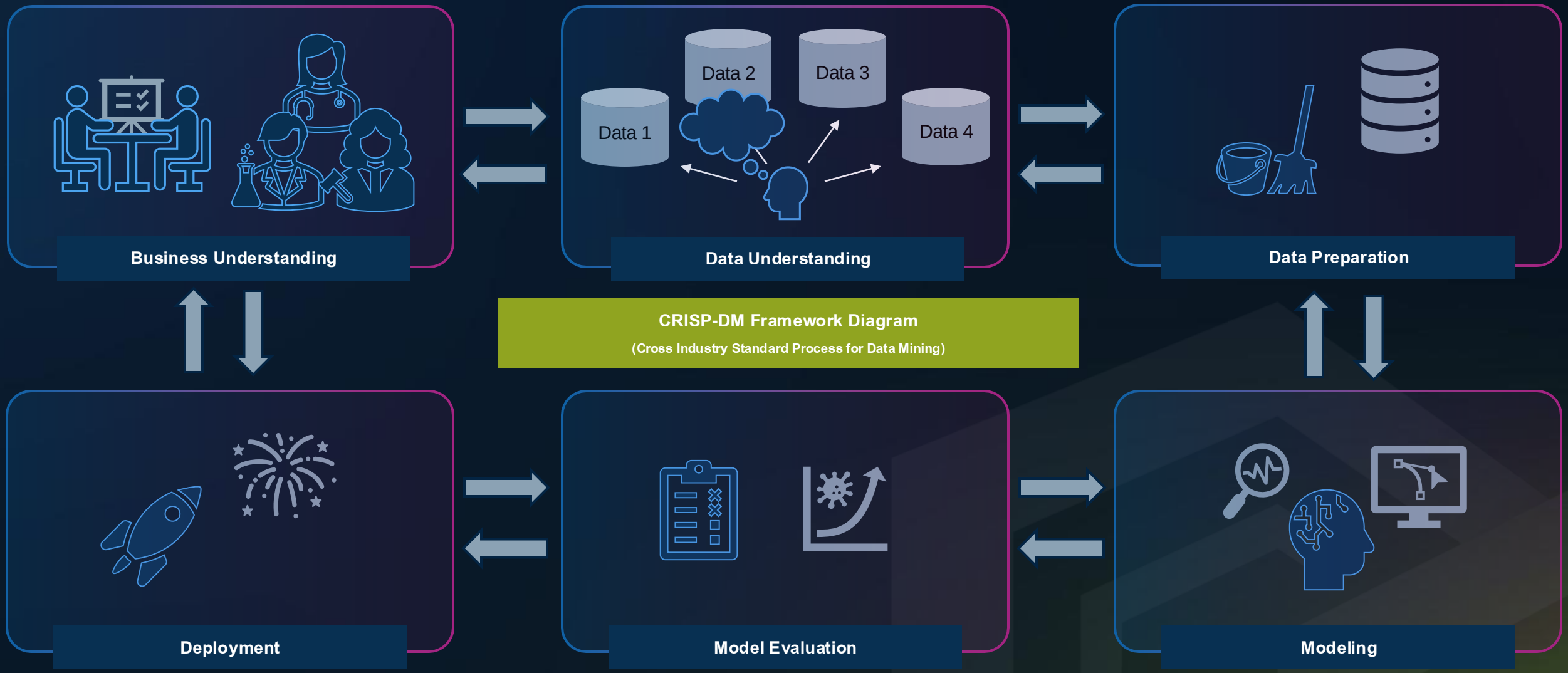
Un-Supervised

X_1	X_2	X_p	Y

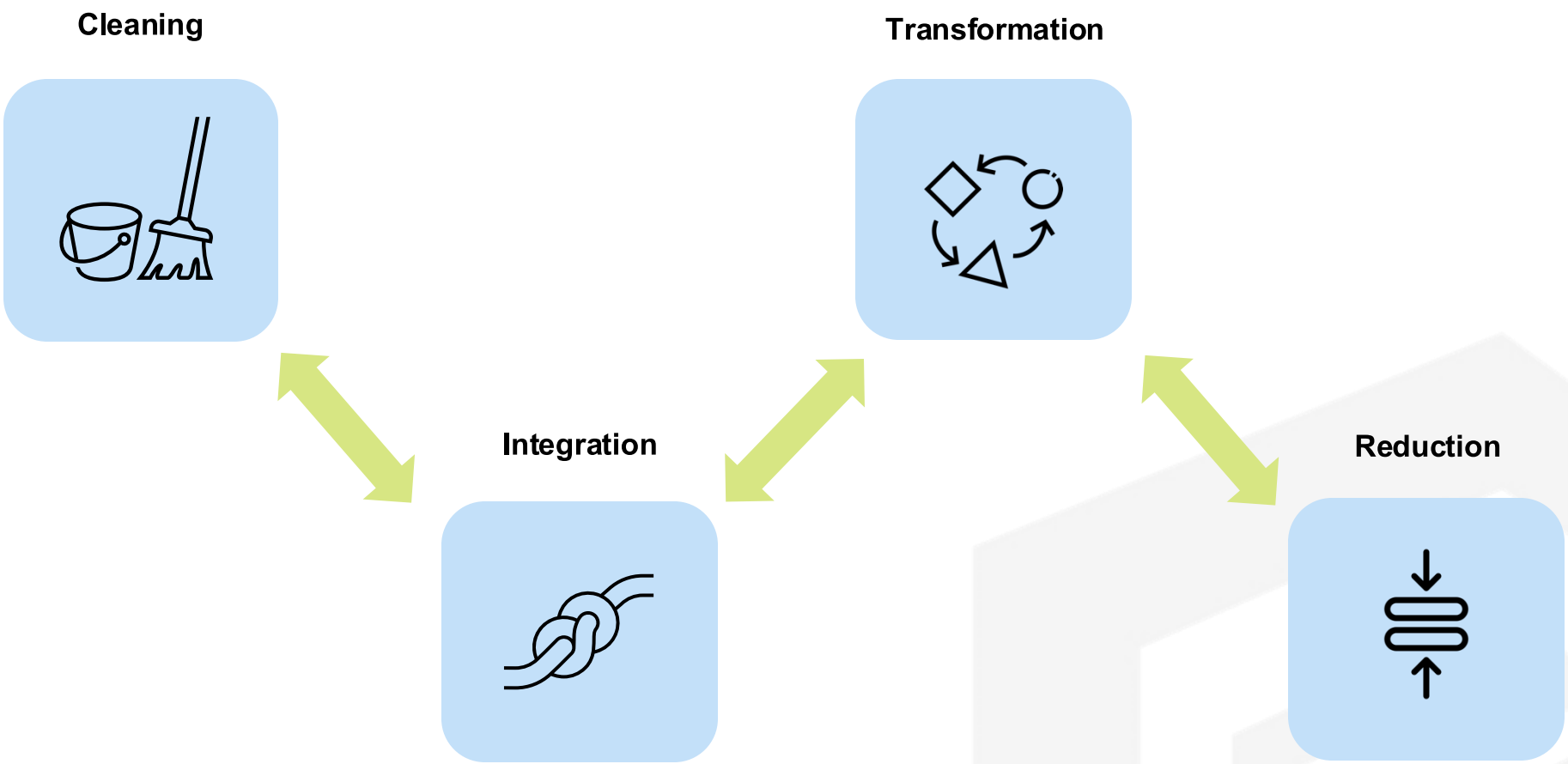
No Target

ML Workflow

STelligence



Data Preprocessing



Data Cleaning

	Item_ID (str)	Item_Name (str)	Category (str)	Quantity (float)	Unit (str)	Last_Updated (date)	Price_THB (str)
1	A001	Rice	Grain	100.0	kg	2024/02/15	1,500
2	A002	Sugar	Grain		kg	15-02-2024	800
3	A003	Flour	Grain	50.0	kg	2024-02-16	400
4	A003	Flour	Grain	50.0	kg	2024-02-16	400
5	A004	Coffe	Beverage	1000.	grams	2024-02-17	12000
6	A005	Tea	Beverage	30.0	kg	2024-02-18	600
7	A006	Oil	Oil	-20.0	liter	2024/02/19	300
8	A007	Rice	grain	100.0	kg	2024-02-20	1,500
9	A008	Salt	Grain	5000.	grams	2024-02-21	40,000
10	A009	Suggar	grain	80.0	kg	2024-02-22	800

1 Missing Values (Null or N/A)

2 Outliers/ Noise

3 Duplicates

4 Data Type Issues

5 Formatting

6 Unit Issues

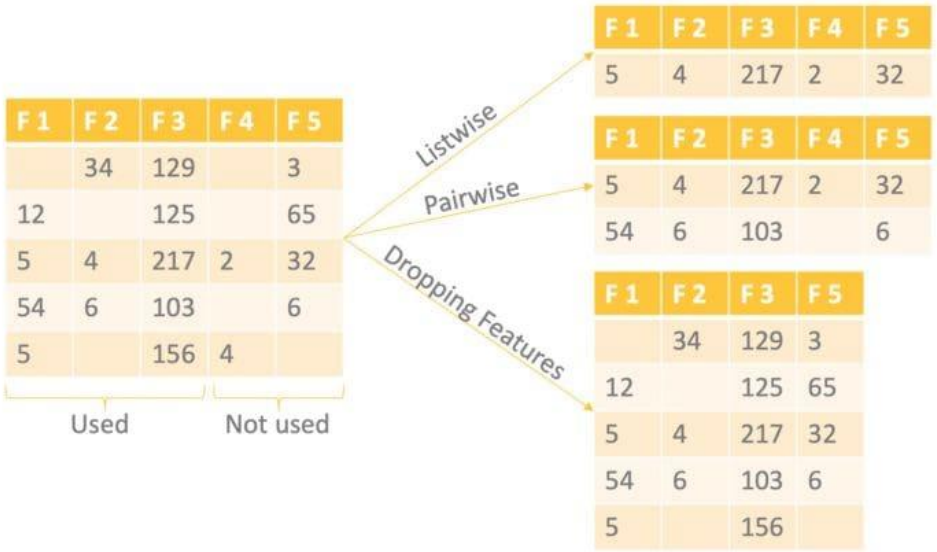
7 Typos/ Synonyms

8 Mislabels

Missing Values

Inspection

- Identify samples containing missing entries.
- Visualize missing data by feature and sample for clear pattern detection.



Source: https://www.kdnuggets.com/2020/09/missing-value-imputation-review.html

Action

- Deletion:
  - Listwise (remove entire samples)
  - Pairwise (use available data in analysis)
- Imputation:
  - Simple methods (mean, median, mode, previous, next)
  - Advanced methods (KNN, regression, deep learning-based methods)
- Interpolation:
  - Apply specifically to sequential or time series data to estimate missing values based on existing trends.

Missing Values

date	price	number_of_news	
1/3/2025	18	8	← 14
2/3/2025	18	8	← 12
3/3/2025	20.1	15	
4/3/2025	20.3	13	
5/3/2025	20.4	20	
6/3/2025	20.2	14	
7/3/2025	20.4	12	
8/3/2025	20.4	12	← 35
9/3/2025	20.4	12	← 40
10/3/2025	20.9	43	
11/3/2025	21	50	
12/3/2025	21.4	54	
13/3/2025	21.5	50	
14/3/2025	21.3	70	
15/3/2025	21.5	56	
16/3/2025	21.8	90	

```
df  = df.fillna(method='ffill')
```

Consider each feature individually,

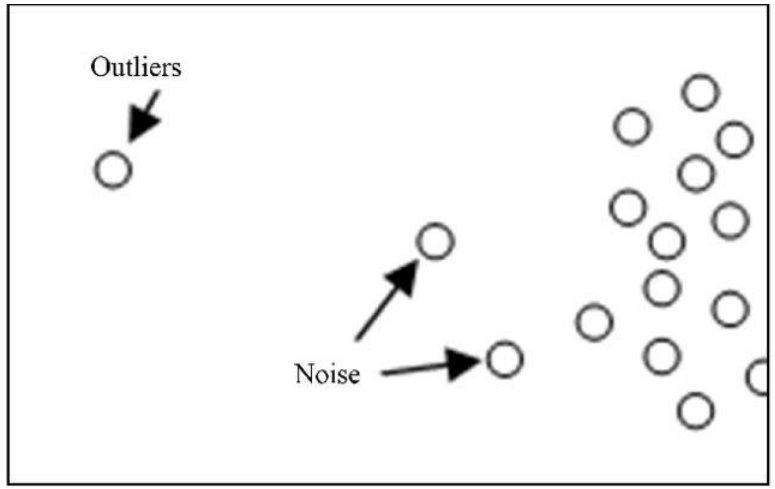
rather than using the same approach for all.

Outliers/ Noise

STelligence

Inspection

- Detect unusual values using statistical measures (Z-score, IQR method, percentile analysis).
- Visualize data distribution clearly with boxplots, histograms, or scatter plots to highlight anomalies.



Action

- Deletion:
  Remove extreme outliers if they're confirmed errors or irrelevant.
- Transformation:
  Use transformations (logarithmic, square-root, Box-Cox) to reduce the effect of outliers.
- Capping/ Winsorizing/ Trimming:
  Set upper/lower bounds based on domain knowledge or statistical thresholds.

“Noise refers to random variations or errors in data, while outliers are data points that deviate significantly from the norm.”

Duplicates

STelligence

Inspection

- Data Entry: Manual input errors or multiple submissions.
- Integration Issues: Duplicate records from merging different datasets.
- Join Operations: Incorrect database joins (inner/outer joins) causing unintended duplications.

Action

- Deletion:
  Eliminate exact duplicates to maintain data integrity.
- Consolidation:
  Merge duplicate entries containing complementary data.
- Prevention:
  Implement unique identifiers or indexing.

Duplicates

id	name	date
493	Ming	7/3/2025
298	Leo	7/3/2025
560	Pan	7/3/2025
330	Gui	7/3/2025
493	Ming	14/3/2025
298	Leo	14/3/2025
560	Pan	14/3/2025
493	Ming	17/3/2025
493	Ming	18/3/2025

```
df  = df.drop_duplicates()
```

Always check what causes duplicated data.

Often, the record is not unique because not all relevant columns are selected.

Data Type Issues

STelligence

Inspection

- Data entry errors (e.g., numeric fields containing special characters).
- File import issues (e.g., CSV/excel imports interpreting numbers as text).
- Inconsistent formatting across different data sources.

Action

- Conversion:
  Explicitly convert data to correct types (numeric, categorical, datetime).
- Cleaning:
  Remove or correct invalid characters or formatting causing incorrect data types (e.g., remove commas from numeric values).

type	*set of values*	*common operators*	*sample literal values*
`int`	integers	`+ - * / %`	`99 12 2147483647`
`double`	floating-point numbers	`+ - * /`	`3.14 2.5 6.022e23`
`boolean`	boolean values	`&& || !`	`true false`
`char`	characters		`'A' '1' '%' '\n'`
`String`	sequences of characters	`+`	`"AB" "Hello" "2.5"`

Formatting

STelligence

Inspection

- Identify inconsistencies in data representation (e.g., different date formats, inconsistent capitalization).
- Check for variations in measurement units, text casing, special characters, or spacing errors.

Date
6/3/21
2021-11-24
11/24/21
5 October 2021
10/5/21
5/12/21
3/28/21
3/28/21
28/3/2021
September 9 2021

Timezone

"Asia/Bangkok"

Action

- Standardization:
  - Convert date formats, units, and numeric separators to a consistent standard.
  - Apply uniform text formatting (e.g., lowercase vs. uppercase).
- Cleaning:
  - Remove unnecessary spaces, special characters, or formatting artifacts (e.g., different currency symbols or thousands separators).
- Validation:
  - Ensure adherence to predefined formatting rules to maintain consistency.

Unit Issues

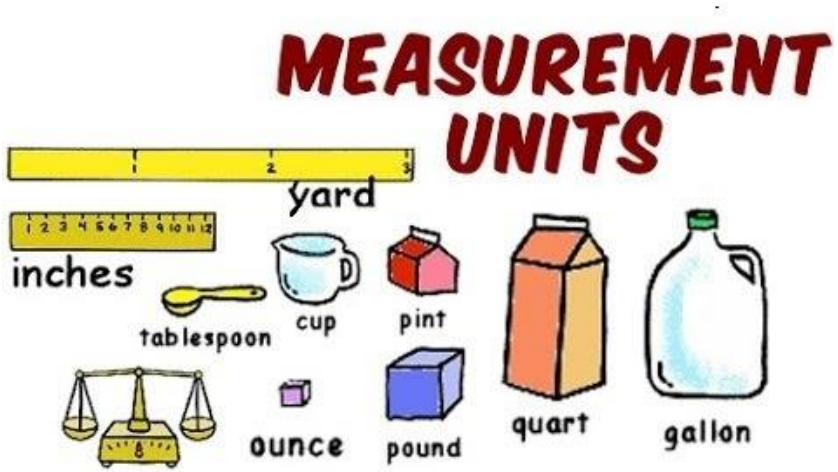
STelligence

Inspection

- Identify inconsistencies in measurement units (e.g., kg vs. grams, liters vs. milliliters).
- Check if numerical values align with the expected unit of measurement.

Action

- Standardization:
  Convert all values to a single, consistent unit based on domain requirements.
- Normalization:
  Apply conversion factors to unify different units (e.g., converting all weight measurements to kilograms).
- Validation:
  Ensure each feature has a predefined unit and enforce consistency across records.



Currency Converter

1 USD = 33.71 THB

Typos/ Synonyms

STelligence

Inspection

- Identify spelling errors, inconsistent capitalization, and variations in word forms (e.g., "Coffe" vs. "Coffee").
- Detect synonyms or alternative terms referring to the same entity (e.g., "Grain" vs. "grain", "Flour" vs. "Wheat Powder").
- Check for extra/missing spaces or special characters causing discrepancies.

Action

- Standardization:
  Create a reference list or dictionary to map all variations to a single standard term.
- Correction:
  Use spell-checking, fuzzy matching, or NLP techniques to detect and fix typos.
- Normalization:
  Convert text to lowercase and remove extra spaces or special characters for uniformity.

สิงห์ คอร์ปอเรชั่น

สิงห์ คอร์ปอเรชัน

แบทเตอรี่

นางสาว

แบตเตอรี

นส.

แบต

สิงห์ คอร์โปเรชั่น

อุปกรณ์เก็บประจุไฟฟ้า

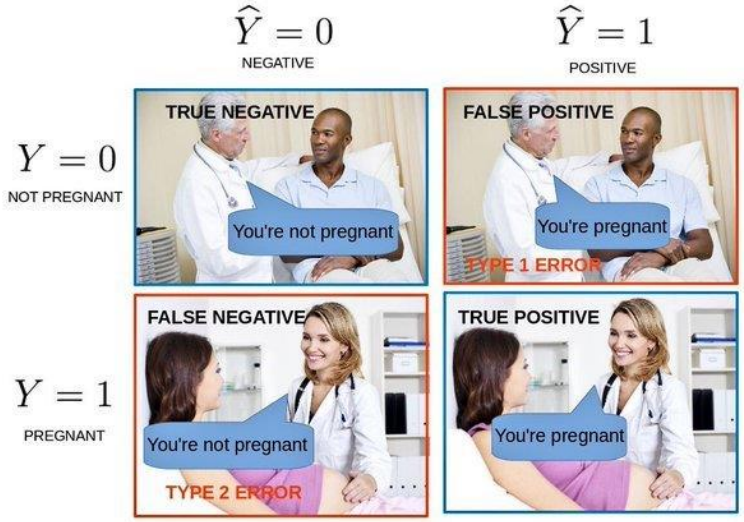
น.ส.

สิงห์ คอร์เปอเรชั่น

Mislabels

Inspection

- Human errors in manual labeling.
- Ambiguous classification where items fit multiple categories.
- Inconsistent taxonomy across different sources or teams.



Action

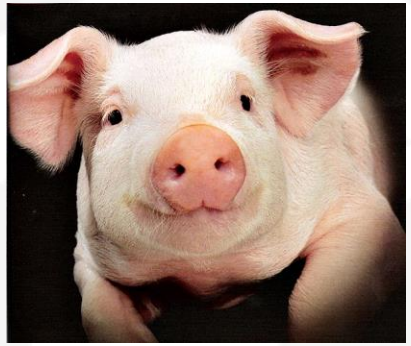
- Standardization:

  Define and enforce a clear category system with predefined valid labels.
- Correction:

  Use rule-based logic, expert validation, or machine learning classification to reassign incorrect labels.
- Auditing:

  Regularly review and verify labels, especially after data integration or transformation.

Label: หมี



Data Integration

STelligence

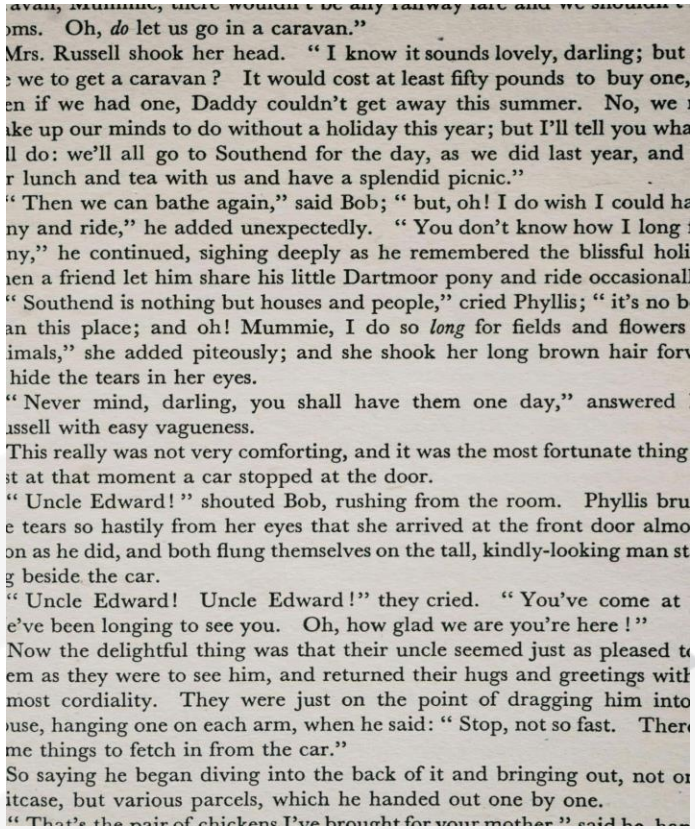
Structured

	A	B	C	D	E
1	First Name	Last Name	Phone Number		
2	John	Doe	555-123-4567		
3	Jane	Smith	555-987-6543		
4	Michael	Johnson	555-345-6789		
5	Emily	Williams	555-876-5432		
6	David	Brown	555-234-5678		
7	Sarah	Taylor	555-654-3210		
8	Christopher	Lee	555-789-0123		
9	Amanda	Miller	555-456-7890		
10	Matthew	Davis	555-890-1234		
11	Jennifer	Wilson	555-321-0987		
12					
13					
14					
15					
16					

Semi-Structured


```
[
  {
    "username": "john_doe",
    "tags": ["travel", "photography"],
    "followers": 1200,
    "following": 345,
    "posts": [
      {"id": 1, "content": "Exploring the world!", "likes": 56},
      {"id": 2, "content": "Capturing breathtaking landscapes.", "likes": 89}
    ]
  },
  {
    "username": "jane_smith",
    "tags": ["fitness", "food"],
    "followers": 2300,
    "following": 567,
    "posts": [
      {"id": 1, "content": "Healthy living is the key!", "likes": 120}
    ]
  },
  {
    "username": "mike_johnson",
    "tags": ["tech", "gaming"],
    "followers": 1750,
    "following": 432,
    "posts": [
      {"id": 1, "content": "Latest tech gadgets review.", "likes": 340},
      {"id": 2, "content": "Gaming marathon all night!", "likes": 420},
      {"id": 3, "content": "Cooking up delicious recipes.", "likes": 210}
    ]
  }
]
```

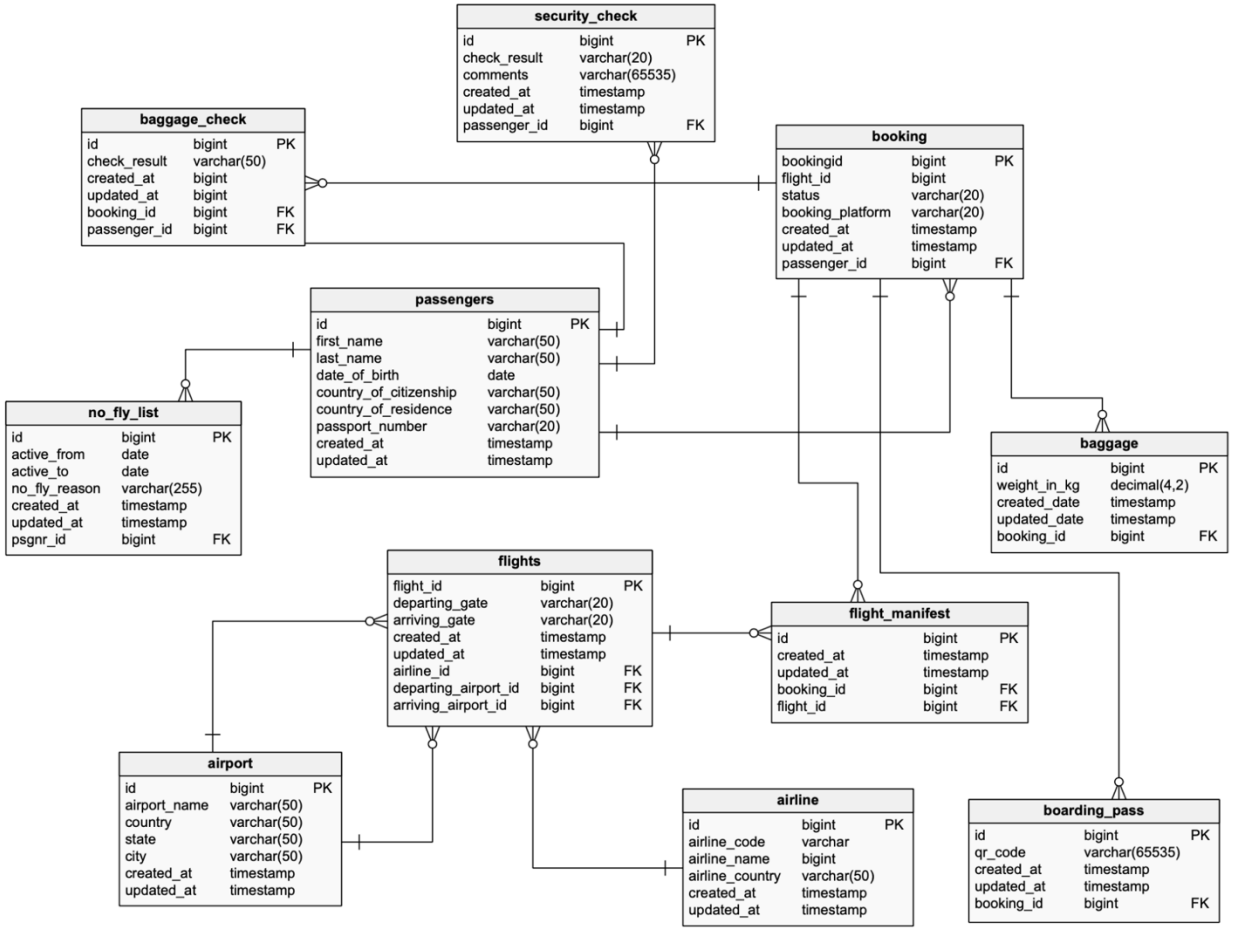

Unstructured



oms. Oh, *do* let us go in a caravan."
Mrs. Russell shook her head. "I know it sounds lovely, darling; but
we to get a caravan? It would cost at least fifty pounds to buy one,
en if we had one, Daddy couldn't get away this summer. No, we
ke up our minds to do without a holiday this year; but I'll tell you wha
ll do: we'll all go to Southend for the day, as we did last year, and
r lunch and tea with us and have a splendid picnic."
"Then we can bathe again," said Bob; "but, oh! I do wish I could ha
ny and ride," he added unexpectedly. "You don't know how I long
ny," he continued, sighing deeply as he remembered the blissful holi
en a friend let him share his little Dartmoor pony and ride occasional
"Southend is nothing but houses and people," cried Phyllis; "it's no b
an this place; and oh! Mummie, I do so *long* for fields and flowers
imals," she added piteously; and she shook her long brown hair for
hide the tears in her eyes.
"Never mind, darling, you shall have them one day," answered
ussell with easy vagueness.
This really was not very comforting, and it was the most fortunate thing
st at that moment a car stopped at the door.
"Uncle Edward!" shouted Bob, rushing from the room. Phyllis bru
e tears so hastily from her eyes that she arrived at the front door almo
on as he did, and both flung themselves on the tall, kindly-looking man st
g beside the car.
"Uncle Edward! Uncle Edward!" they cried. "You've come at
e've been longing to see you. Oh, how glad we are you're here!"
Now the delightful thing was that their uncle seemed just as pleased t
em as they were to see him, and returned their hugs and greetings with
most cordiality. They were just on the point of dragging him into
ouse, hanging one on each arm, when he said: "Stop, not so fast. There
me things to fetch in from the car."
So saying he began diving into the back of it and bringing out, not o
itcase, but various parcels, which he handed out one by one.

Source: https://blog.datath.com/structured-unstructured-semistructured-data/

Data Integration

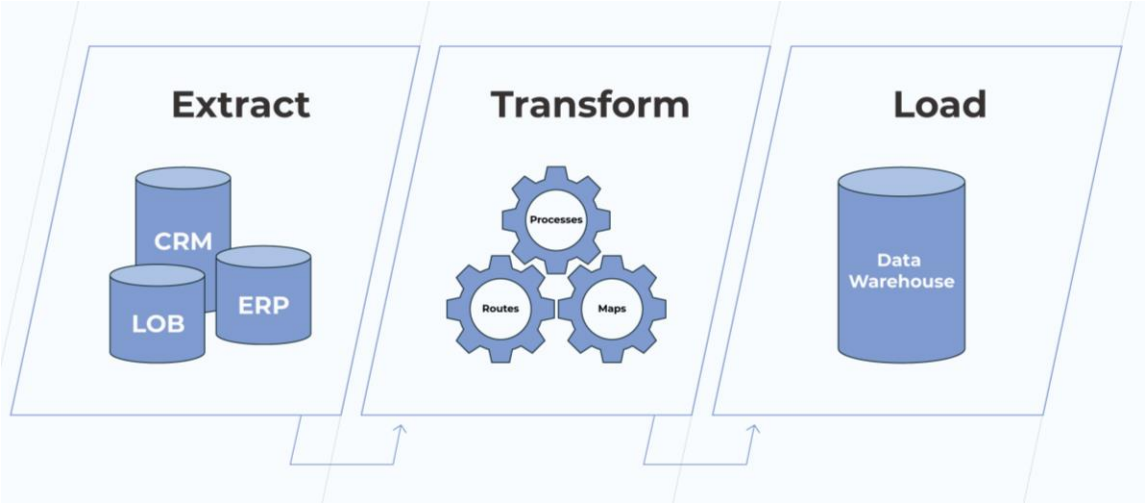


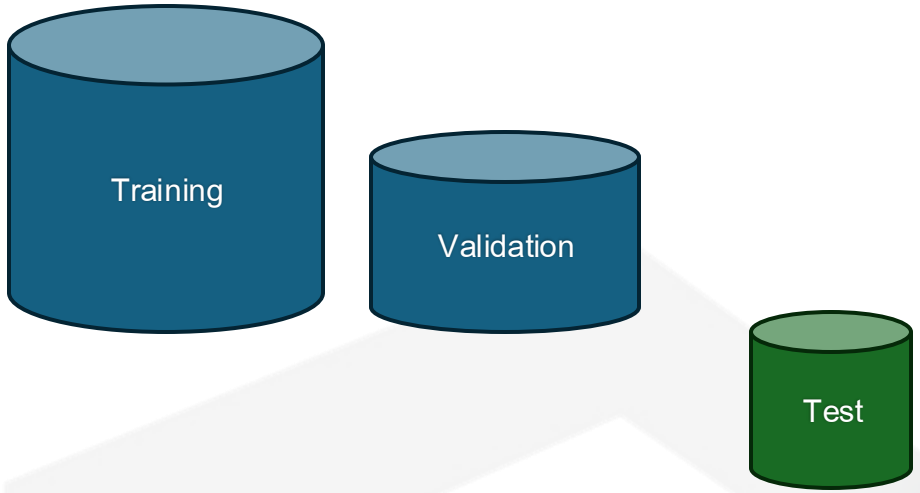
Source: https://vertabelo.com/blog/vertabelo-tips-good-er-diagram-layout/

String Case Type

- Camel Case → myVariableName
- Pascal Case → MyVariableName
- Snake Case → my_variable_name
- Kebab Case → my-variable-name
- Title Case → My Variable Name

Data Integration



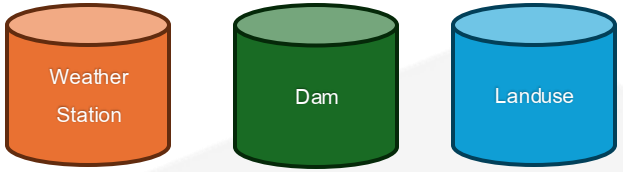


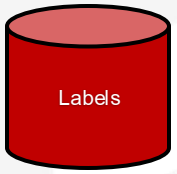
Data Transformation

Clearly define the samples based on the business goal, data sources, and modeling techniques.

Source: https://kids.nationalgeographic.com/science/article/flood

I want to obtain flood predictions for the next 3 days
in each province of Thailand.





Labels

- Flash Flood (น้ำท่วมฉับพลัน)
- Waterlogging (น้ำท่วมขัง)
- Overflow (น้ำล้นตลิ่ง)

Data Transformation

Explore “Labels” first among all data sources

Labels

- Daily data from 2020/12/06 to 2024/12/31
- Subdistrict (ตำบล)
- Classes = [“Flash Flood”, “Waterlogging”, “Overflow”]

Weather Station

- Daily data from 2021/12/01 to 2024/12/31
- Station (Latitude, Longitude)

Dam

- Weekly data from 2021/12/01 to 2023/12/31
- Dam (Latitude, Longitude)

Landuse

- Monthly data from 2019/12/01 to 2024/12/31
- Province (จังหวัด)

I want to obtain flood predictions for the next 3 days in each province of Thailand.

Daily data from … to …

Province

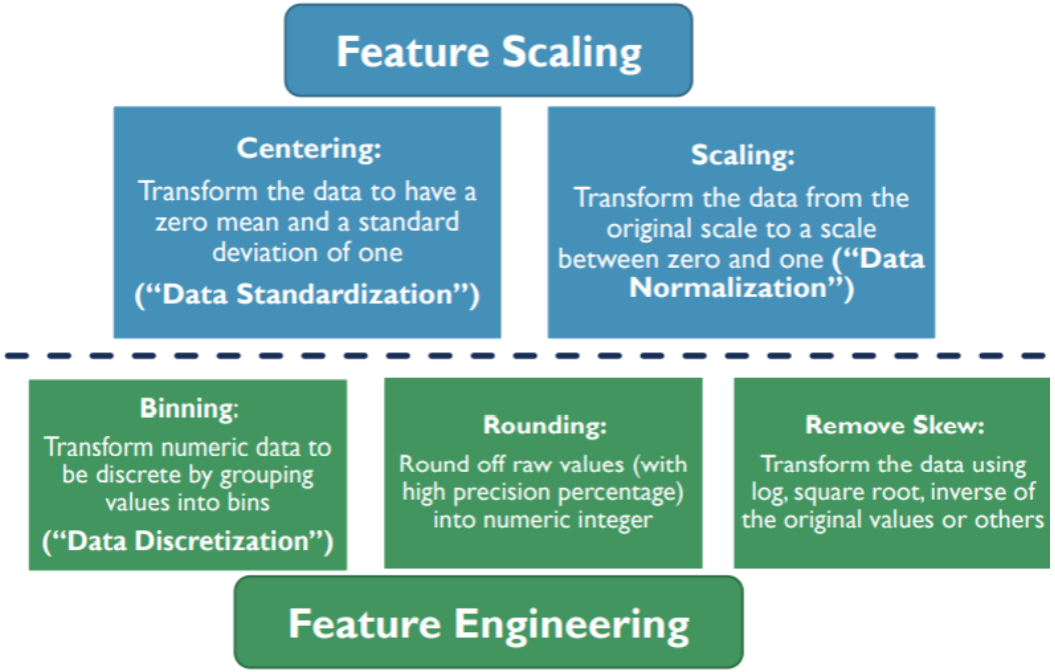
Flood (Y=1), Not (Y=0)

Accuracy = 0.5 !!!

Go back to CRISP-DM. Let’s try another way.

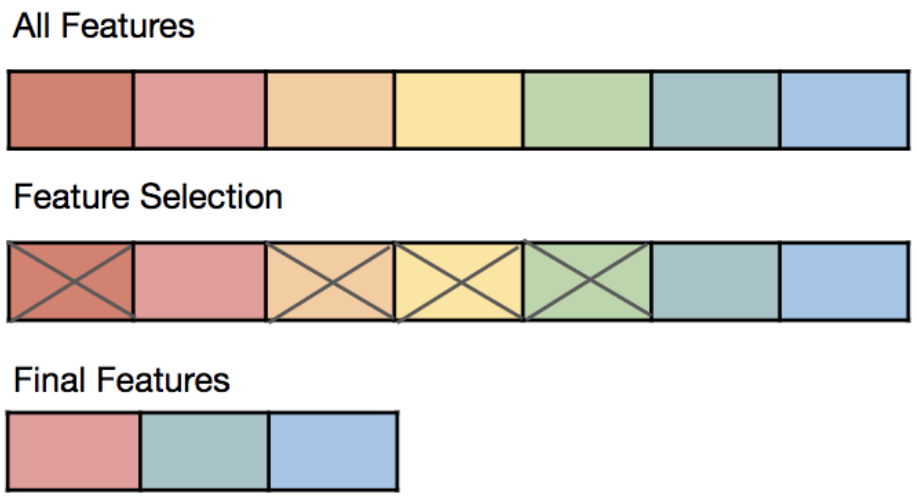
Data Transformation

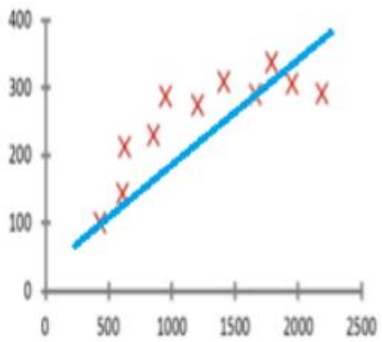
Feature Engineering



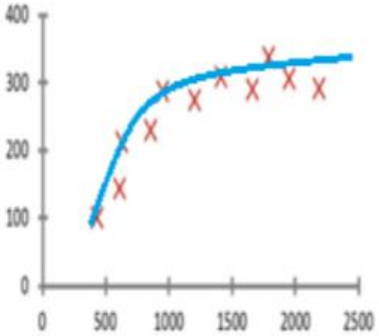
- Feature Creation
- Feature Transformation
- Feature Extraction
- Feature Selection

Data Reduction

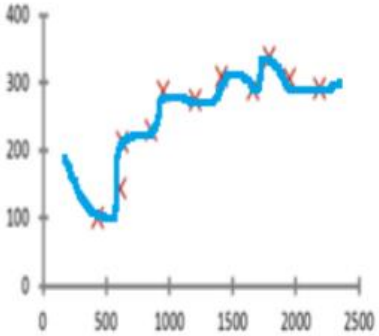




Underfit: "High bias"

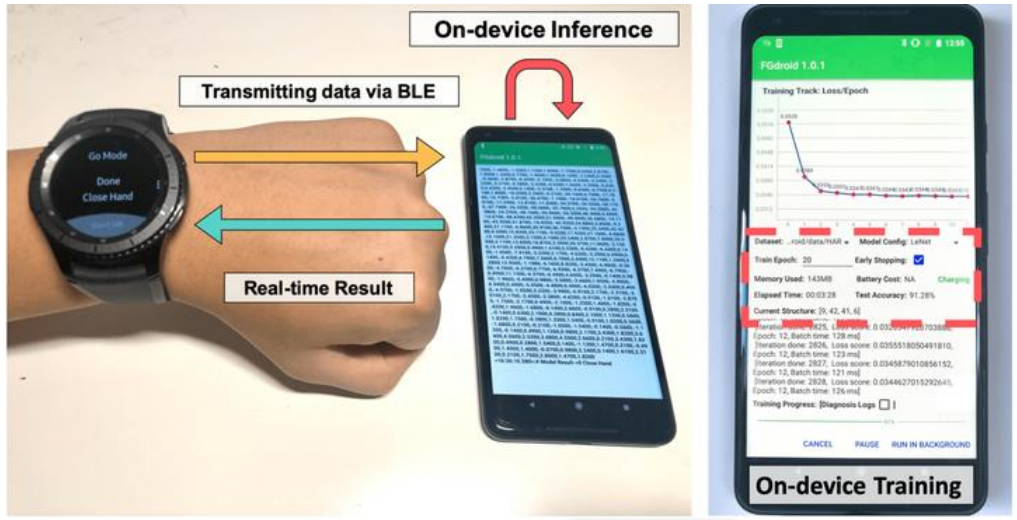


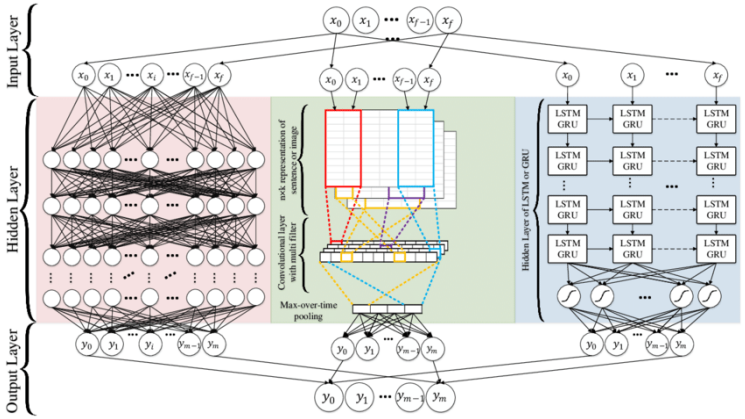
Good balance:
"Low bias, Low variance"



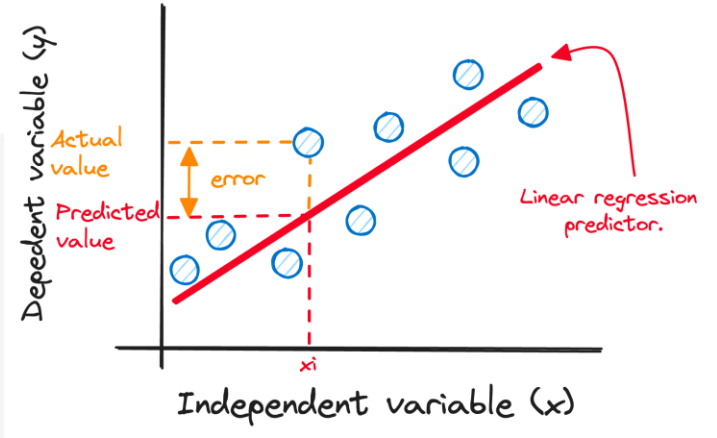
Overfit: "High variance"

Data Reduction





vs



Data Reduction

STelligence

Model A

vs

Model B

- Data requires: 2,000,000 samples
- Training time: 5 hours
- Accuracy: 0.99

Good for research

or when high precision is required.

- Data requires: 2,000 samples
- Training time: 5 mins
- Accuracy: 0.90

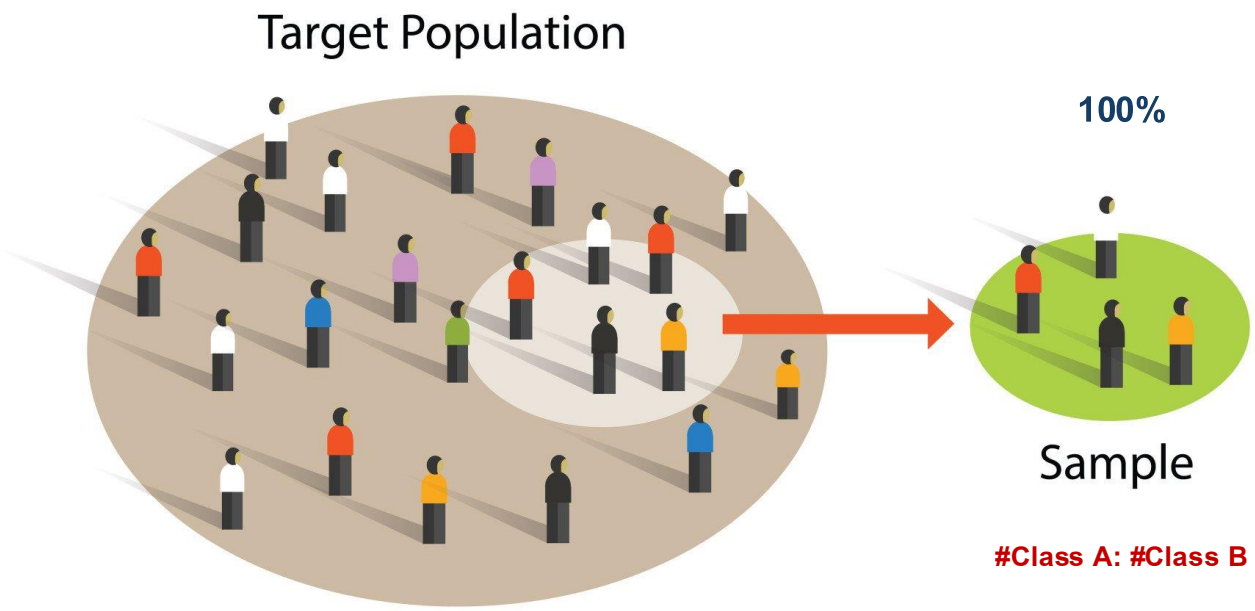
Good for business

where retraining is needed.

Data Reduction

Data Reduction

Stratified **Sampling**

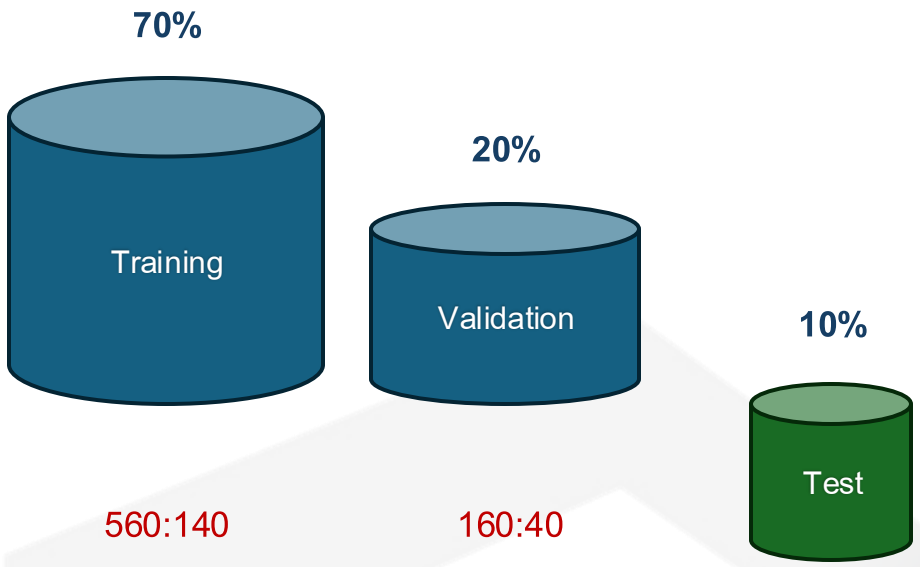


#Class A: #Class B

800:200

4,000:1,000

Binary Classification (n = 1,000)



560:140

160:40

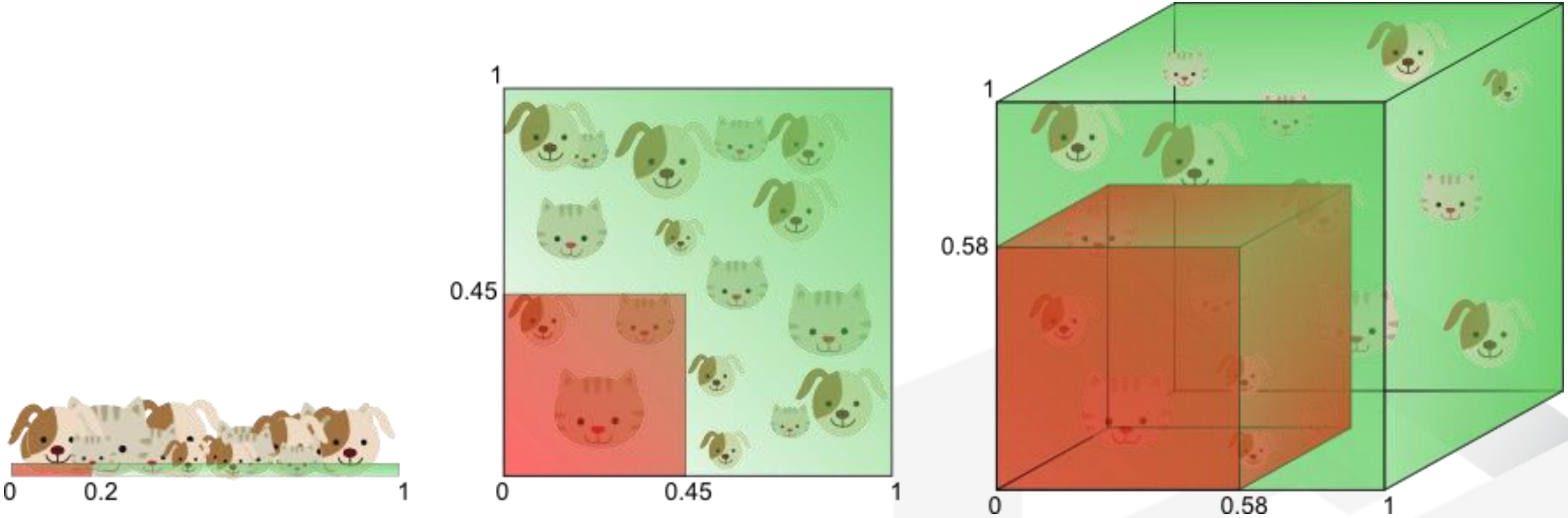
80:20

Random repetitions with different seeds!

Source: https://www.simplypsychology.org/sampling.html

Data Reduction

Curse of Dimensionality



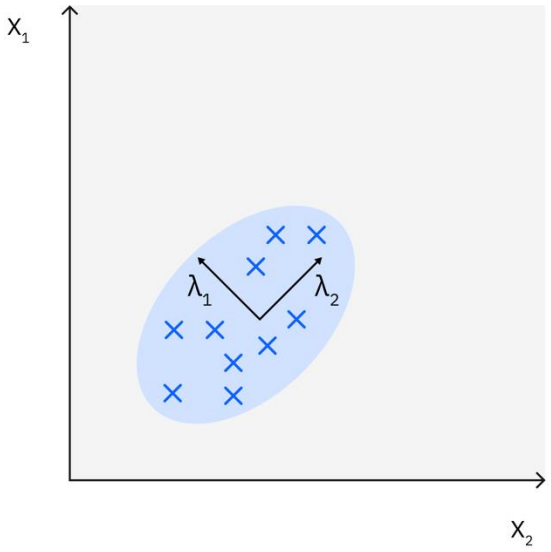
Source: https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

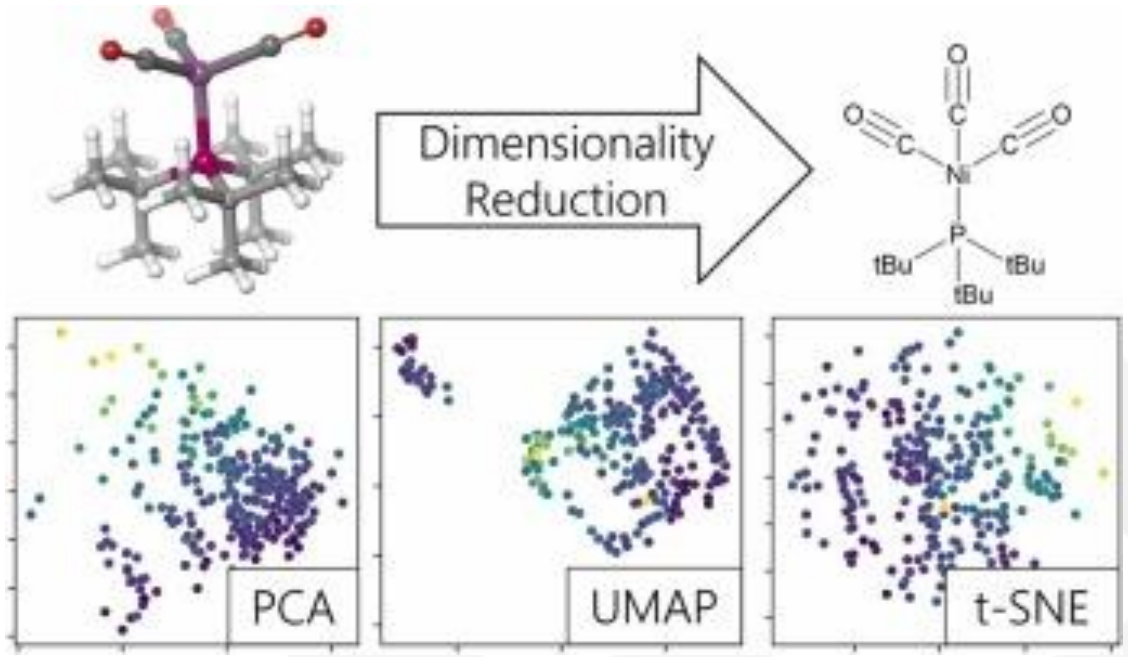
Data Reduction

Principal component analysis (PCA)

Projections that maximize variance = eigenvectors (PC)

PCA:





Source: https://www.sciencedirect.com/science/article/pii/S2949747724000137

STelligence

Break 15 minutes

Group Assignment (Score 20 points):

- **Each group must consist of 3-5 members**
- **Task Requirement (15 points):**

  Using the provided dataset, find at least **10** insights based on the following criteria:
  - Insights from a single table → 0.5 points
  - Insights from 2 tables → 1 point
  - Insights from more than 2 tables → 2 points
- **Submission Requirement (5 points):**

  Each group must submit the list of group members, and the scripts used to derive each insight via **myCourseVille**
- **Bonus Points:**

  Groups that can find unique insights different from others will receive **extra points**
- **Example Insight:**

  Why does Product "A" have high sales in March? → Evidence: Use graphs or data to support the conclusion
- **Provided Dataset:**

  The dataset comes from a retail company and consists of 4 tables and 2 master files:

  1. **sales.csv** ⟶ Sales data of products at each branch
  2. **order.csv** ⟶ Order data of products at each branch
  3. **receive.csv** ⟶ Data on received products at each branch
  4. **inventory.xlsx** ⟶ Inventory data at each branch
  5. **master_sku.xlsx** ⟶ Product information
  6. **master_site.xlsx** ⟶ Store branch information

Note: You may use Open Data from APIs or perform Web Scraping to enhance the analysis.

Data Dictionary

Sales table:

Field Name	Type	Description
trans_date	DATE	Transaction date of the sale
site_no	VARCHAR	Unique identifier for the site or store
sku_no	VARCHAR	Stock Keeping Unit (SKU) number identifying the product
inv_no	VARCHAR	Invoice number associated with the transaction
customer_id	VARCHAR	Unique identifier for the customer
sale_qty	DECIMAL	Quantity of items sold in the transaction
salev_amt	DECIMAL	Total sales amount before discounts
disc_amt	DECIMAL	Discount amount applied to the transaction
promo_type	VARCHAR	Type of promotion applied to the sale

Order table:

Field Name	Type	Description
date_order	DATE	Date when the order was placed
ord_no	VARCHAR	Unique order number
ord_issue_by	VARCHAR	Identifier of the person or system issuing the order
site_no	VARCHAR	Unique identifier for the site or store
sku_no	VARCHAR	Stock Keeping Unit (SKU) number identifying the product
order_qty	DECIMAL	Quantity of items ordered
orderv_amt	DECIMAL	Total order value before any adjustments

Data Dictionary

Receive table:

Field Name	Type	Description
date_rcv	DATE	Date when the goods were received
site_no	VARCHAR	Unique identifier for the site or store
sku_no	VARCHAR	Stock Keeping Unit (SKU) number identifying the product
receive_vamt	DECIMAL	Total value of received goods (before adjustments)
receive_qty	DECIMAL	Quantity of items received

Inventory table:

Field Name	Type	Description
date_stk	DATE	Date of stock measurement
site_no	VARCHAR	Unique identifier for the site or store
sku_no	VARCHAR	Stock Keeping Unit (SKU) number identifying the product
end_amt	DECIMAL	Total inventory value at the end of the period
end_qty	DECIMAL	Total quantity of items in stock at the end of the period

Data Dictionary

Master SKU table:

Field Name	Type	Description
sku_no	VARCHAR	Stock Keeping Unit (SKU) number identifying the product
class_name	VARCHAR	Classification name of the product
subclass_name	VARCHAR	Subclassification name for more detailed grouping

Master Site table:

Field Name	Type	Description
site_no	VARCHAR	Unique identifier for the site or store
site_name	VARCHAR	Name of the site or store location
lat	DECIMAL	Latitude coordinate of the site location
long	DECIMAL	Longitude coordinate of the site location

STelligence

THANK YOU