



INTRODUCTION TO DATA ANALYTICS

Data Processing

↓
เรื่อง เชื้อไวรัสโควิด

↓
เพื่อเป็นข้อมูลขององค์กร

Dr. Rathachai Chawuthai

Department of Computer Engineering

Faculty of Engineering

King Mongkut's Institute of Technology Ladkrabang

Agenda

- Data Processing
- Data Cleansing
- Data Integration
- Data Transformation
- Feature Engineering

Data Processing



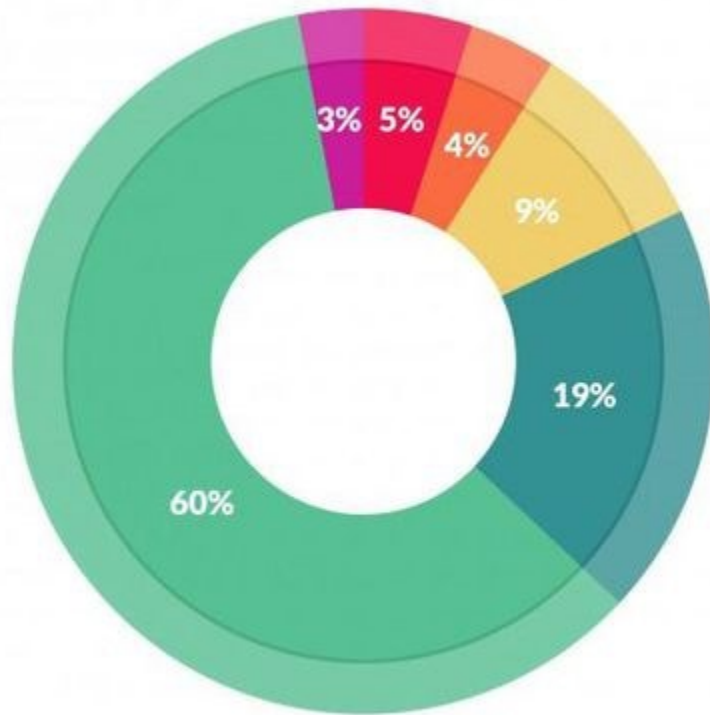
Is Data Beautiful? → data ទំនើប មិនសូវ, មិនសូវល្អ (ត្រូវការការសម្រួល)
ទាំងអស់ (clean)



Why Is Data Dirty? → data မြေရှင်းရန်

- Incomplete data comes from → data မြေရှင်းရန် → မြေရှင်းရန်
 - n/a data value when collected
 - different consideration between the time when the data was collected and when it is analyzed.
 - human/hardware/software problems
- Noisy data comes from the process of data → noisy data
 - collection
 - entry
 - transmission
- Inconsistent data comes from → data မြေရှင်းရန် → မြေရှင်းရန်
 - Different data sources
 - Functional dependency violation

Spending the most time doing ...



What data scientists spend the most time doing

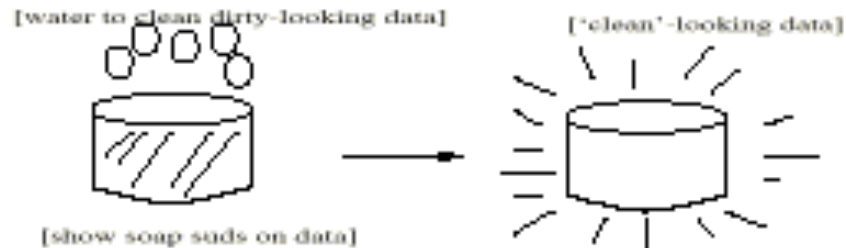
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Tasks

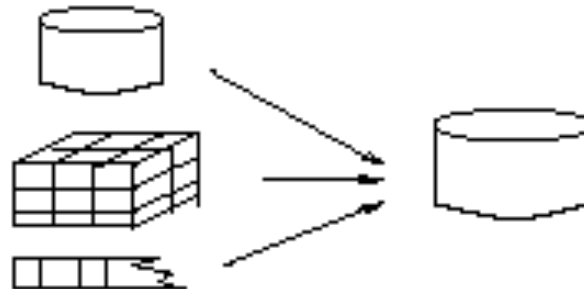
- Data Cleaning ការកាត់ចោលនិងការកែសម្រួល
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data Integration ការរួបរួមទិន្នន័យ → ឯកតាមការប្រើប្រាស់
 - Integration of multiple databases, data cubes, or files
- Data Transformation ការប្រែប្រួលទិន្នន័យ
 - Normalization and aggregation
- Data Reduction ការកាត់បន្ថយទិន្នន័យ (NEXT TIME) (ឯកតាមការប្រើប្រាស់)
 - Obtains reduced representation in volume but produces the same or similar analytical results

Data Processing

Data Cleaning



Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



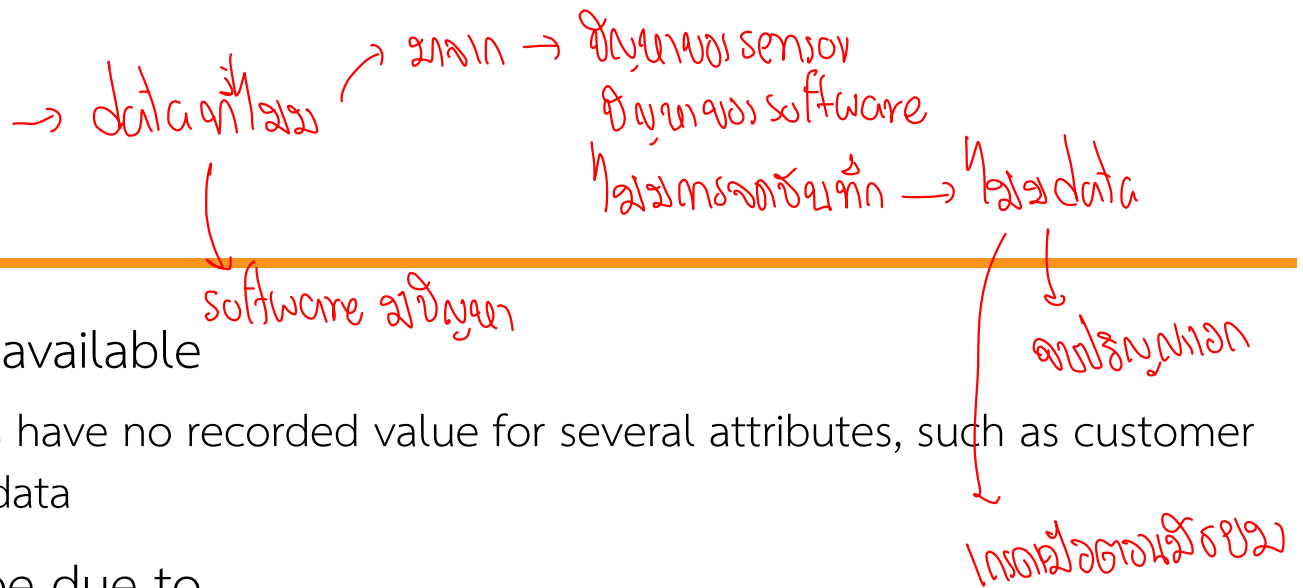
Data Cleansing



Data Cleaning

- Importance
 - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
 - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks → *งานทำความสะอาด*
 - Fill in missing values → *ใส่ข้อมูลที่หายไป*
 - Identify outliers and smooth out noisy data → *หา outliers และ smooth* (ขจัดค่าผิดปกติ)
 - Correct inconsistent data → *แก้ไข data ที่ขัดแย้งกัน*
 - Resolve redundancy caused by data integration → *ลบ data ที่ซ้ำกัน*

Missing Data



- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - equipment malfunction
 - inconsistent with other recorded data and thus deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - not register history or changes of the data
- Missing data may need to be inferred.

infer

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
 - a global constant : e.g., “unknown”, a new class?! → *new class for missing values*
 - the attribute mean *ค่าเฉลี่ย*
 - the attribute mean for all samples belonging to the same class: smarter
↓ *ค่าเฉลี่ยของแต่ละ class*
 - the most probable value: inference-based such as Bayesian formula or decision tree
inference algorithm

Options for Dealing With Missing Data

Missing data in general is one of the trickier issues that is dealt with when cleaning data. Broadly there are two solutions:

- Deleting/Ignoring rows with missing values → *lösungs → Lösungen*
- Filling in the Values → *erfüllen*

Deleting/Ignoring rows with missing values

↪ လေ့လာမှု လုပ်ငန်းဆောင်ရွက်ရာတွင် null တွေကို ဖယ်ရှားပစ်ရမည် → ဒီလိုလုပ်ရတာမှန်ပဲ

- The **simplest solution** available when faced with missing values is to **not use the records with missing values** when training your model. However, there are some issues to be aware of before you starting deleting masses of rows from your dataset.
- The first is that this approach only makes sense **if the number of rows with missing data is relatively small** compared to the dataset. If you are finding that you will be deleting more than around 10% of your dataset due to rows having missing values, you may need to reconsider.

Filling in the Values

benjamin

- The second broad option for dealing with missing data is to fill the missing values with a value. But what value to use? This depends on a range of factors, including the type of data you are trying to fill.
- If the data is categorical (i.e. countries, device types, etc.), it may make sense to simply create a new category that will represent 'unknown'.
class → unknown
- Another option may be to fill the values with the most common value for that column (the mode). However, because these are broad methods for filling the missing values, this may oversimplify your data and/or make your final model less accurate.
a common 91 1850 mode

Filling in the Values

↑ ໄຂ້ດວງເຕັມຕົວທ່າປ → ຢາດເວັ້ນຈື່ງຂຽນຕົວຈິງ
ສຳຫລັບການ ປຸກ mean ຂອງ class ຈື່ງ → ເພດຊາຍ ເພດຍິງ

- For numerical values (for example the age column) there are some other options. Given that in this case using the mode to fill values makes less sense, we could instead use the mean or median. We could even take an **average based on some other criteria** – for example filling the missing age values based on **an average age for users that selected the same country_destination**.
- For both types of data (categorical and numerical), we can also use far more complicated methods to impute the missing values. Effectively, we can use a similar methodology that we are planning to use to predict the country_destination to **predict the values** in any of the other columns, based on the columns that do have data. And just like with modeling in general, there are an almost endless number of ways this can be done, which won't be detailed here.

Educated Guessing

- It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4.

	Strongly Disagree 1	2	3	4	Strongly Agree 5
The Chipotle website has a clean and simple presentation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
The Chipotle website is easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
I find the Chipotle website to be attractive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
The Chipotle website is trustworthy.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
It is easy to navigate within the Chipotle website.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The Chipotle website's capabilities meet my requirements.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
The information on the Chipotle website is credible.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
I will likely return to the Chipotle website in the future.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

ทั้งหมดนี้
น่าจะตอบ
4 เหมือนกัน

Noisy Data → ខ្សែអន្តរាគមន៍ → ឥទ្ធិពល ៣០០ ២០១៧ បណ្តាញ ↓ ឆ្លងកាត់

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- Other data problems which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

លក្ខណៈ ធម៌ ឧបករណ៍

វិធីសាស្ត្រ → ចំនួន $Q_1 - Q_3$

ចំនួន = $Q_1 - 1.5 IQR$

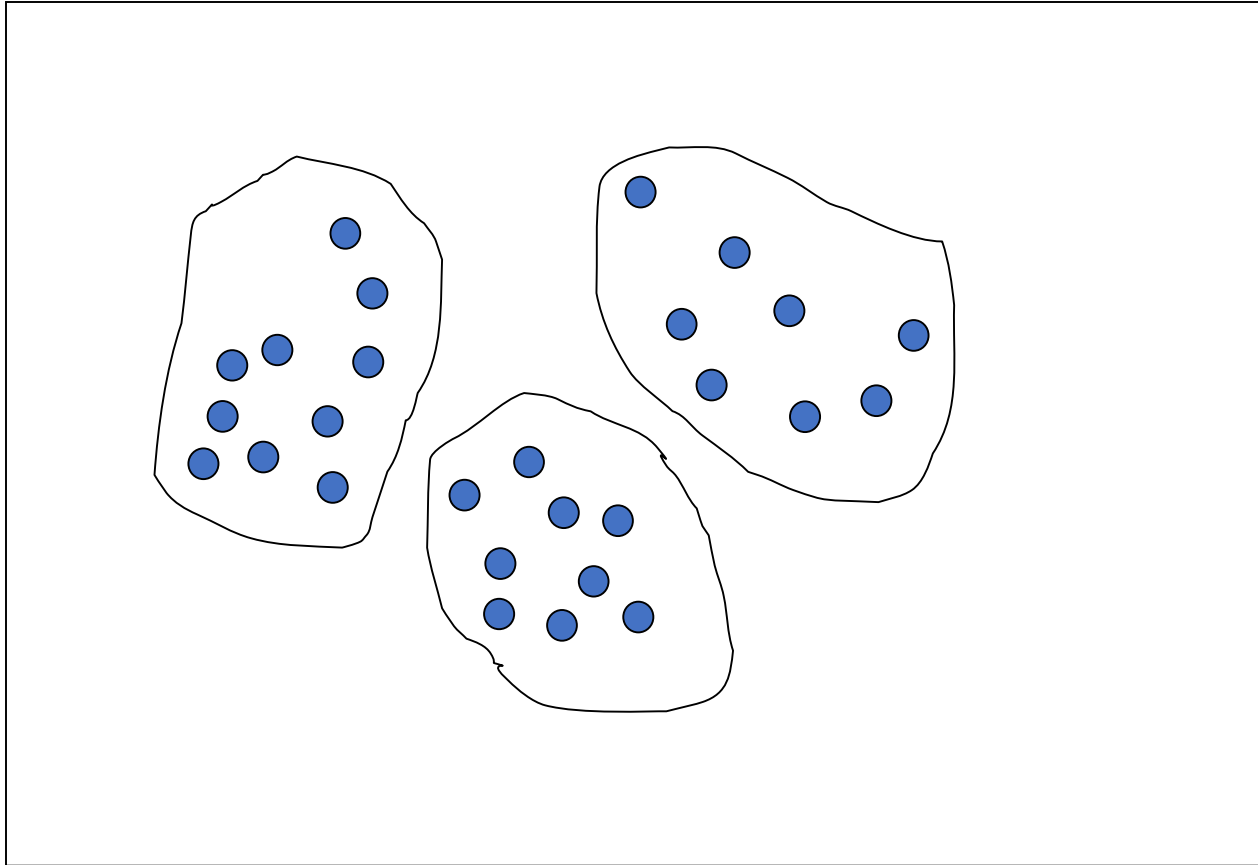
ចំនួន = $Q_3 + 1.5 IQR$

ឧបករណ៍ = noise

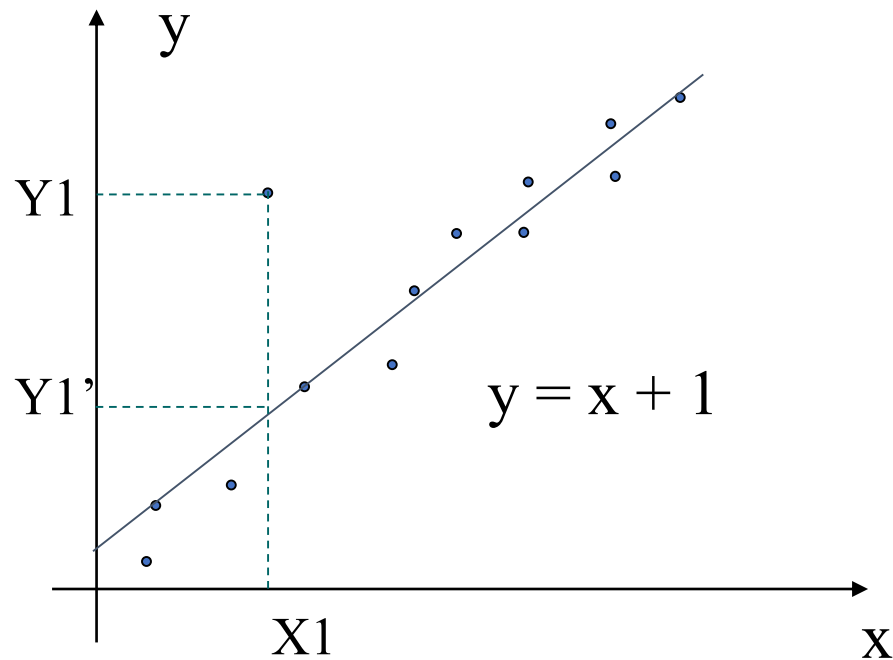
How to Handle Noisy Data? → จัดการไว้

- Clustering
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human (e.g., deal with possible outliers)
- Regression
 - smooth by fitting the data into regression functions

Cluster Analysis



Regression



Data Integration



Data Integration

ប្រធាន

- Data integration: → រួបរួម data ចេញពី source ជាច្រើន format, ទុំ តែមួយ
 - combines data from multiple sources into a coherent store
- Schema integration → ប្រើ schema ចំណុះ → ប្រើ ថ្នាក់ column
 - integrate metadata from different sources
 - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id \equiv B.cust-#
- Detecting and resolving data value conflicts → data មាន conflict
 - for the same real world entity, attribute values from different sources are different
 - possible reasons: different representations, different scales, e.g., metric vs. British units

Data Integration

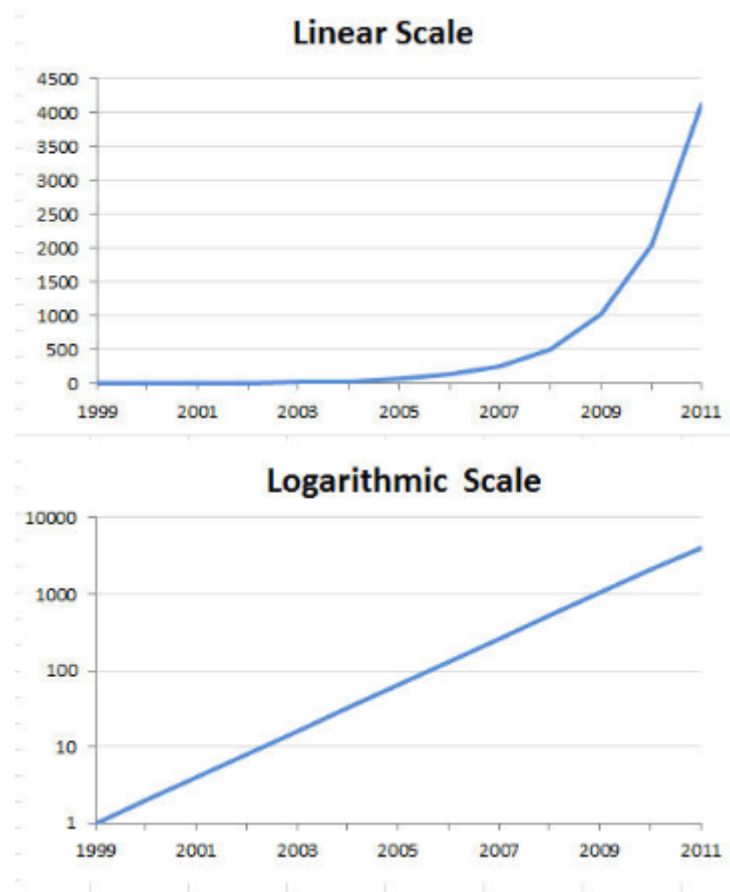


Data Transformation

မလေ့လာရသေးသော data

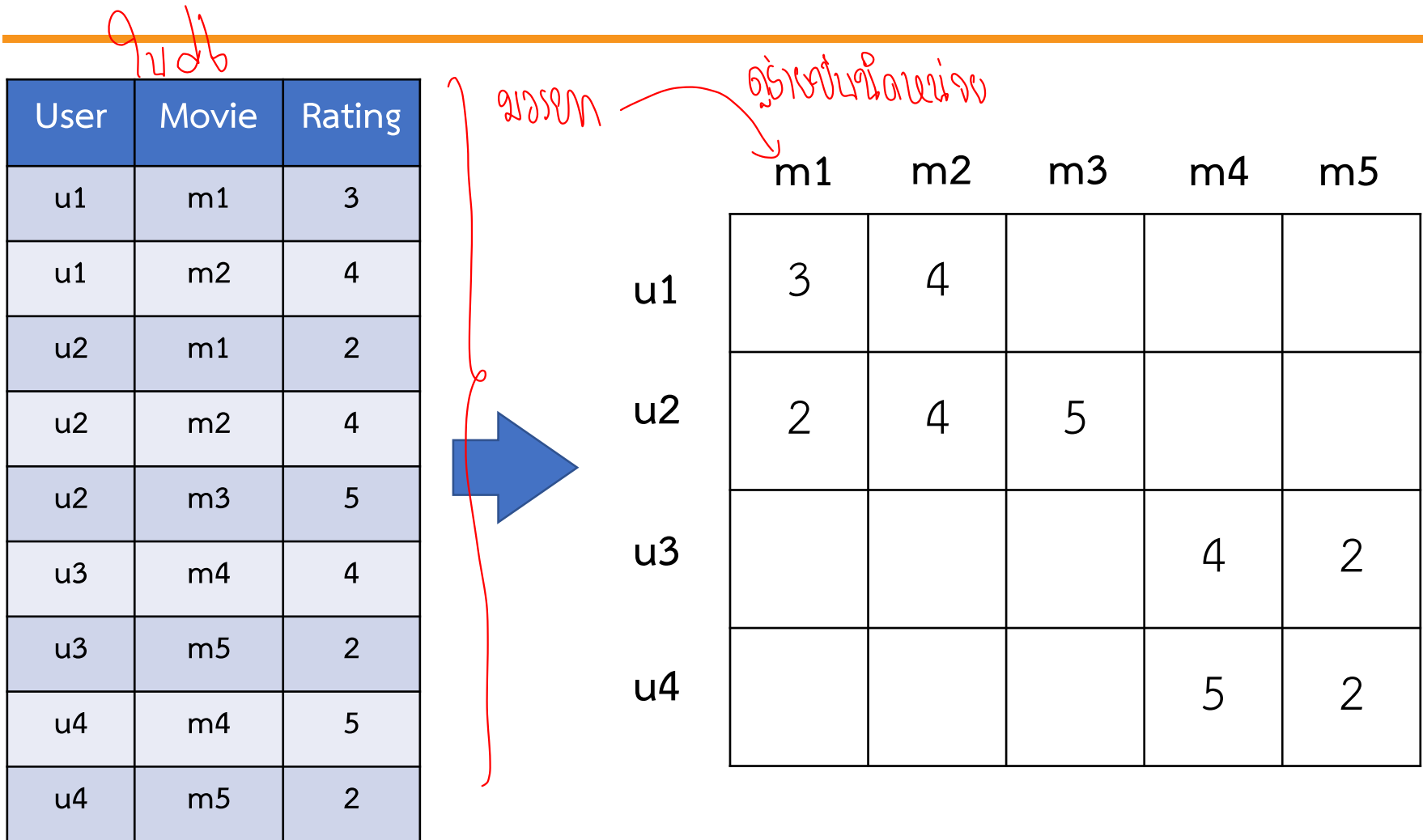
- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
- Attribute/feature construction
 - New attributes constructed from the given ones

Reshape



$\log_e(x)$ → ขวบนปรกนเปนก๊อง
เลขนต
แล้ว1000?

Sparse Matrix

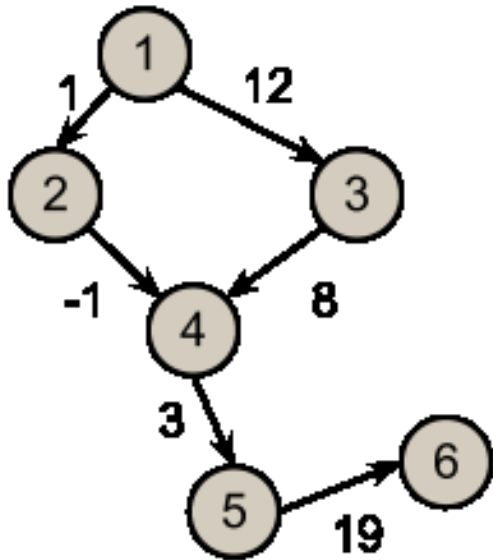


Document to Vector → *အကယ်၍ပါမကလေးသော doc ကို*

ဖော်

	<i>D1</i>	<i>D2</i>	<i>D3</i>	<i>D4</i>	<i>D5</i>	<i>...</i>
<i>a</i>	145	223	346	78	89	<i>...</i>
<i>abandon</i>	4	0	0	5	3	<i>...</i>
<i>ability</i>	5	10	0	4	7	<i>...</i>
<i>able</i>	31	35	64	3	5	<i>...</i>
<i>about</i>	64	68	89	24	9	<i>...</i>
<i>above</i>	4	5	8	0	0	<i>...</i>
<i>abroad</i>	0	0	1	0	0	<i>...</i>
<i>absence</i>	2	4	0	0	0	<i>...</i>
<i>absent</i>	0	0	1	0	0	<i>...</i>
<i>absolute</i>	3	1	5	0	1	<i>...</i>
<i>abstract</i>	5	1	2	1	0	<i>...</i>
<i>abuse</i>	0	1	0	0	0	<i>...</i>
<i>academic</i>	1	3	0	0	0	<i>...</i>
<i>...</i>	<i>...</i>	<i>...</i>	<i>...</i>	<i>...</i>	<i>...</i>	<i>...</i>

Adjacency Matrix



Weighted Directed Graph

→ ગીચગાંઠ & સ્વચ્છ

	1	2	3	4	5	6
1	0	1	12	0	0	0
2	-1	0	0	-1	0	0
3	-12	0	0	8	0	0
4	0	1	-8	0	3	0
5	0	0	0	-3	0	19
6	0	0	0	0	-19	0

Adjacency Matrix

* Feature Engineering



Feature Engineering → ការប្រែប្រួល feature ឱ្យបានល្អ

ក្នុងនោះមាន៖

- Imputation
- Handling Outliers
- Binning
- Log Transform
- One-Hot Encoding
- Grouping Operations
- Feature Split
- Scaling
- Extracting Date

Imputation

→ missing data → null (imputation) → class → mode / class
→ num → mean / median

- Filling missing values
- Numerical Imputation
 - Replace with 0 is not a good solution
 - Mean may be better
 - Mean of the group
 - Same in cluster
 - Regression
- Categorical Imputation
 - Replacing the missing values with the **maximum occurred value**.
 - If values are distributed uniformly and there is not a dominant value, imputing a category like “**Other**” might be more sensible, because in such a case, your imputation is likely to converge a random selection.

Handling Outliers → ตรวจจับค่าผิดปกติ → interquartile

↳ { $\bar{x} + sd$, $\bar{x} - sd$ } จะค้นหาค่าผิดปกติ

- These will detect them using standard deviation, and percentiles.
- Outlier Detection with Standard Deviation
 - If a value has a distance to the average higher than $x * \text{standard deviation}$, it can be assumed as an outlier. Then what x should be?
- Outlier Detection with Percentiles
 - To assume a certain percent of the value from the top or the bottom as an outlier. The key point is here to set the percentage value once again, and this depends on the distribution of your data.
 - Detect data between 5th and 95th percentiles.
 - An Outlier Dilemma: Drop or Cap

Binning

កម្រិតព័ត៌មានទទួលបាន → បាត់បង់ព័ត៌មានកម្រិតទទួលបាន/ឱ្យប្រព័ន្ធគេរ → naive baye

- The main motivation of binning is to make the model more robust and prevent overfitting, however, it has a cost to the performance. Every time you bin something, you sacrifice information and make your data more regularized.

ឱ្យប្រព័ន្ធគេរ

- ## Numerical Binning Example

<u>Value</u>	<u>Bin</u>
0-30 ->	Low
31-70 ->	Mid
71-100 ->	High

បាត់បង់ព័ត៌មាន → កម្រិតព័ត៌មានទទួលបាន → ឱ្យ knowledge ប្រើប្រាស់បាន

class

- ## Categorical Binning Example

<u>Value</u>	<u>Bin</u>
Spain ->	Europe
Italy ->	Europe
Chile ->	South America
Brazil ->	South America

↓
បាត់បង់ព័ត៌មាន

Log Transform \rightarrow ប្រែប្រួលទិន្នន័យ (log) ឱ្យមាន លក្ខណៈ ធម្មតា

- Logarithm transformation (or log transform) is one of the most commonly used mathematical transformations in feature engineering. What are the benefits of log transform:
 - It helps to handle skewed data and after transformation, the distribution becomes more approximate to normal.
 - In most of the cases the magnitude order of the data changes within the range of the data. For instance, the difference between ages 15 and 20 is not equal to the ages 65 and 70. In terms of years, 5 years of difference in young ages mean a higher magnitude difference. This type of data comes from a multiplicative process and log transform normalizes the magnitude differences like that.
 - It also decreases the effect of the outliers, due to the normalization of magnitude differences and the model become more robust.

One-Hot Encoding → 1 column → string values → multiple columns (1 column per value)

- One-hot encoding is one of the most common encoding methods in machine learning. This method spreads the values in a column to multiple **flag columns** and **assigns 0 or 1** to them. These binary values express the relationship between grouped and encoded column.
- This method changes your categorical data, which is challenging to understand for algorithms, to a numerical format and enables you to group your categorical data without losing any information. (For details please see the last part of Categorical Column Grouping)

One-Hot Encoding

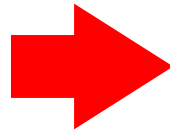
User	City
1	Roma
2	Madrid
1	Madrid
3	Istanbul
2	Istanbul
1	Istanbul
1	Roma



User	Istanbul	Madrid
1	0	0
2	0	1
1	0	1
3	1	0
2	1	0
1	1	0
1	0	0

One-Hot Encoding

User	Social Network
A	Facebook, Twitter
B	Instagram
C	Instagram, Facebook



User	Facebook	Instagram	Twitter
A	1	0	1
B	0	1	0
C	1	1	0

code in python

```
encoded_columns = pd.get_dummies(data['column'])  
data = data.join(encoded_columns).drop('column', axis=1)
```

จางชื่อ column drop รั้งนี้

Grouping Operations

- feature vector = ឈ្មោះ + ប្រាក់*ទំហំ
is represented by a
how a different feature
ទៅបំពេញវ៉ិចទ័របាន

Grouping Operations

- Categorical Column Grouping

User	City	Visit Days
1	Roma	1
2	Madrid	2
1	Madrid	1
3	Istanbul	1
2	Istanbul	4
1	Istanbul	3
1	Roma	3



User	Istanbul	Madrid	Roma
1	3	1	4
2	4	2	0
3	1	0	0

- Numerical Column Grouping

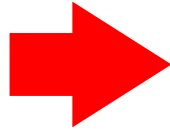
- Numerical columns are **grouped** using **sum** and **mean** functions in most of the cases. Both can be preferable according to the meaning of the feature.

Feature Split *begin: non string → non column*

- Splitting features is a good way to make them useful in terms of machine learning. Most of the time the dataset **contains string columns that violates tidy data** principles. By extracting the utilizable parts of a column into new features:
 - We enable machine learning algorithms to comprehend them.
 - Make possible to bin and group them.
 - Improve model performance by uncovering potential information.
- Split function is a good option, however, there is no one way of splitting features. It depends on the characteristics of the column, how to split it. Let's introduce it with two examples. First, a simple split function for an ordinary name column:

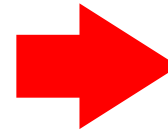
Feature Split

user	name
1	John Doe
2	Peter Smith
3	Tony Brown



user	first_name	last_name
1	John	Doe
2	Peter	Smith
3	Tony	Brown

Movie
Toy Story (1995)
Jumanji (1995)
Grumpier Old Men (1995)



Year
1995
1995
1995

Scaling → data normalization → \$, B

- In most cases, the numerical features of the dataset **do not have a certain range and they differ from each other**. In real life, it is nonsense to expect age and income columns to have the same range. But from the machine learning point of view, how these two columns can be compared?
- Scaling solves this problem. The continuous features become identical in terms of the range, after a scaling process. This process is not mandatory for many algorithms, but it might be still nice to apply. However, **the algorithms based on distance calculations such as k-NN or k-Means need to have scaled continuous features as model input**.

Scaling

normalization → ပုံစံတူညီစေရန် 0-1

min-max normalization

- **Normalization** (or min-max normalization) scale all values in a fixed range between 0 and 1. This transformation does not change the distribution of the feature and due to the decreased standard deviations, the effects of the outliers increases. Therefore, before normalization, it is recommended to handle the outliers.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

value	normalized
2	0.23
45	0.63
-23	0.00
85	1.00
28	0.47
2	0.23
35	0.54
-12	0.10

Scaling

- **Standardization** (or z-score normalization) scales the values while taking into account standard deviation. If the standard deviation of features is different, their range also would differ from each other. **This reduces the effect of the outliers in the features.**
- In the following formula of standardization,
 - the mean is shown as μ
 - the standard deviation is shown as σ .

$$z = \frac{x - \mu}{\sigma}$$

นี่ z-score 11994

↓
จำนวนจริง → จำนวนจริง → จำนวนเต็มลบหรือจำนวนเต็มบวก

value	normalized
2	-0.52
45	0.70
-23	-1.23
85	1.84
28	0.22
2	-0.52
35	-0.42
-12	-0.92

Extracting Date → ทำให้อันที่ซับซ้อน → สามารถสร้าง feature ได้มากมาย

- It might be the reason for this, that dates can be present in numerous formats, which make it **hard to understand by algorithms**, even they are simplified to a format like **"01-01-2017"**.
- Building an ordinal relationship between the values is very challenging for a machine learning algorithm if you leave the date columns without manipulation. Here, three types of preprocessing for dates are suggested:
 - Extracting the **parts of the date into different columns**: Year, month, day, etc.
 - Extracting the **time period between the current date and columns** in terms of years, months, days, etc.
 - Extracting **some specific features from the date**: Name of the weekday, Weekend or not, holiday or not, etc.

Extracting Date

- If you transform the date column into the extracted columns, the information of them become disclosed and machine learning algorithms can easily understand them.

```
from datetime import date

data = pd.DataFrame({'date':
['01-01-2017',
'04-12-2008',
'23-06-1988',
'25-08-1999',
'20-02-1993',
]})
```

Extracting Date

#Transform string to date

`data['date'] = pd.to_datetime(data.date, format="%d-%m-%Y")`

→ convert string

↓

แปลงจาก string pd to datetime

#Extracting Year

`data['year'] = data['date'].dt.year`

#Extracting Month

`data['month'] = data['date'].dt.month`

#Extracting passed years since the date

`data['diff_years'] = date.today().year - data['date'].dt.year`

#Extracting passed months since the date

`data['diff_months'] = (date.today().year - data['date'].dt.year) * 12 + date.today().month - data['date'].dt.month`

#Extracting the weekday name of the date

`data['day_name'] = data['date'].dt.day_name()`

Extracting Date

date	year	month	diff_years	diff_months	day_name
2017-01-01	2017	1	2	26	Sunday
2008-12-04	2008	12	11	123	Thursday
1988-06-23	1988	6	31	369	Thursday
1999-08-25	1999	8	20	235	Wednesday
1993-02-20	1993	2	26	313	Saturday

“

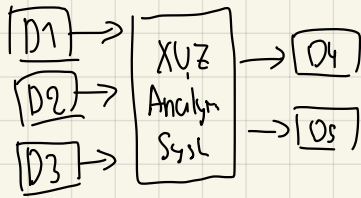
We're entering a new world
in which data may be
more important than software.

”

Tim O'Reilly

Data pipeline

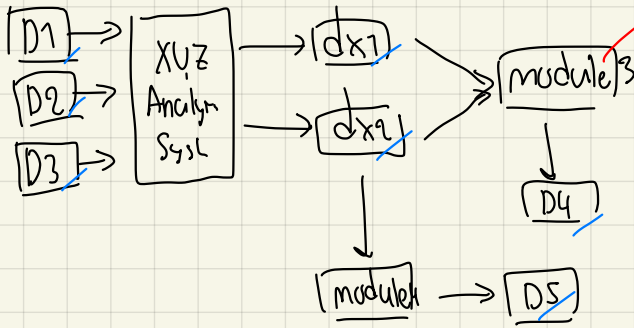
↳ look data flow diagram มาเลย



วอร์ปที่นี่คือให้มันทำอย่างอื่นดูว่าเอา data ไปใช้ที่/บน哪儿

warp here

แต่ถ้าเราทำอย่างอื่นที่มันดูจริงๆ เอาตัวมาเลย



อาจจะเป็น python file

ให้มันทำอย่างอื่นได้
แล้วต้องเก็บของไว้บนเครื่อง

ก็มันเก็บ data ะ
↓
คือ data dic ด้วย

- ① ชื่อของข้อมูล
- ② ถ้ามันเก็บได้
- ③ note: ของมันเก็บ
- ④ schema

field, datatype, des
name string ชื่อ

↳ warp here

รูปแบบสุดท้าย → ตัวของที่เรา input อะไรก็ได้ output อะไรก็ได้

↓
ของมันเป็นอย่างไร → ให้มันทำ