Preprocessing

Data Mining

Man-Kwan Shan

CS, NCCU

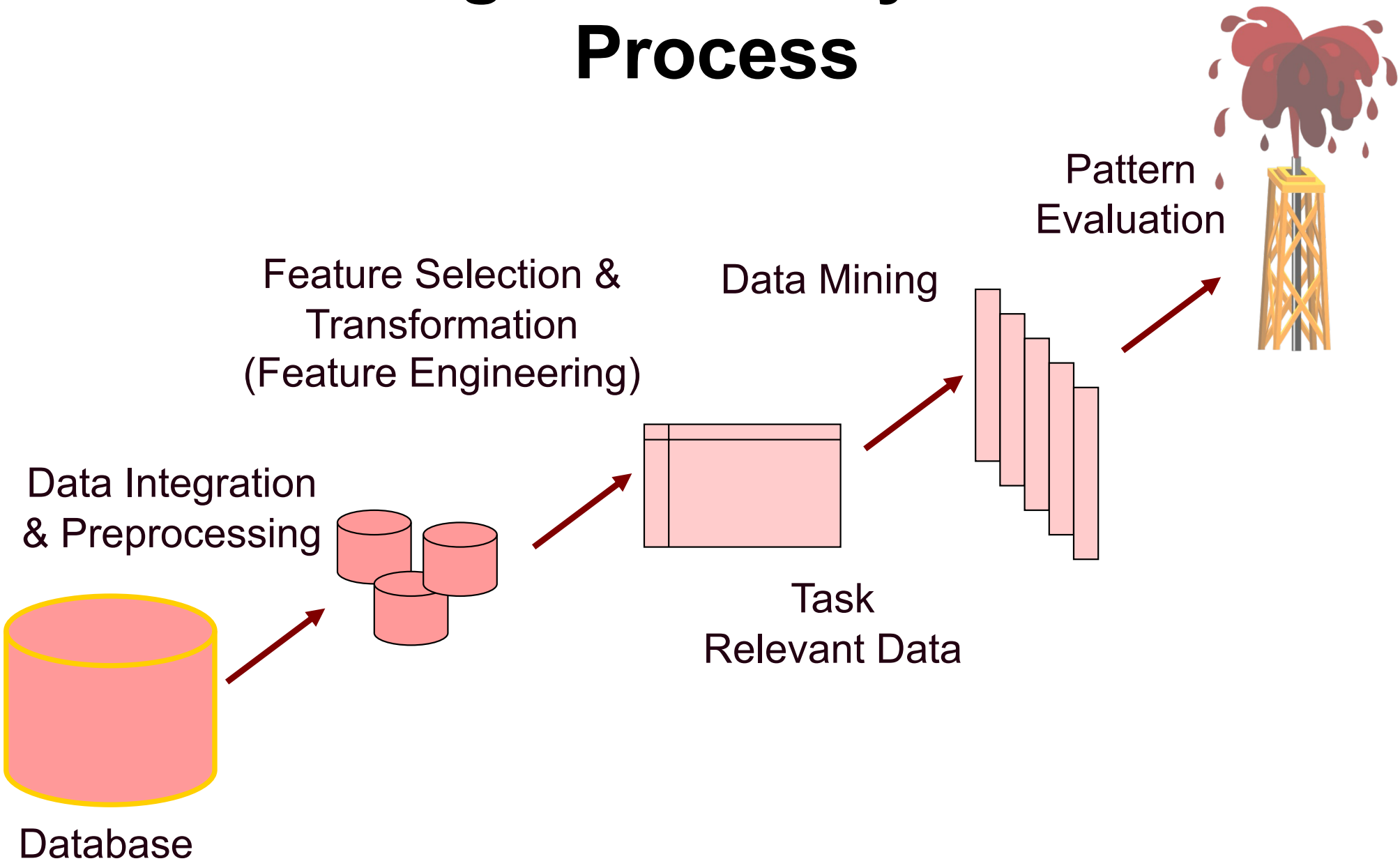
What is Data Mining ?

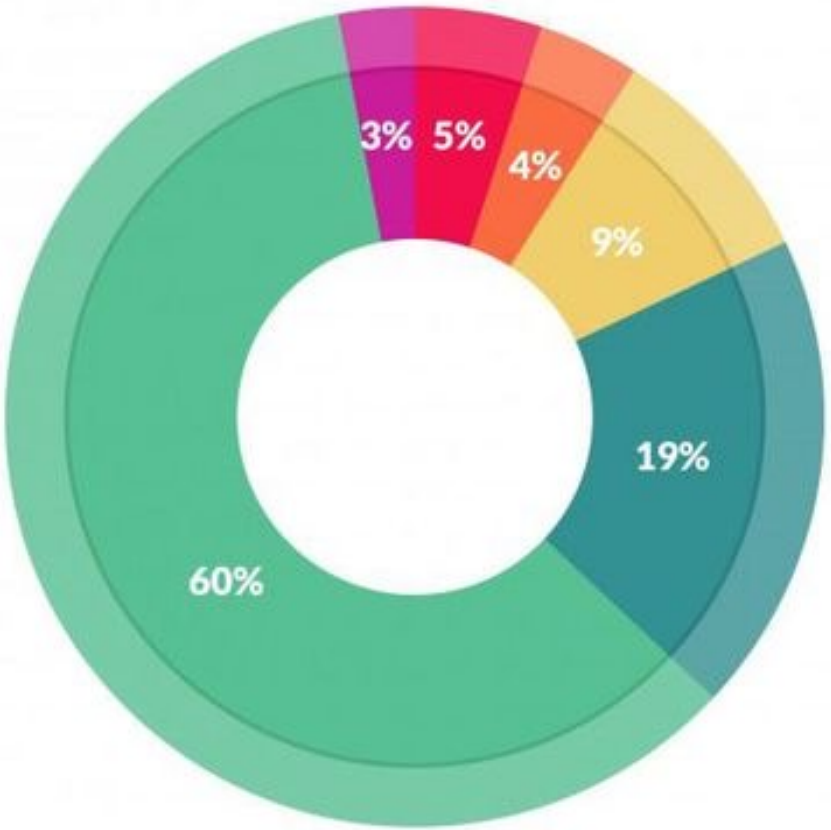
- Nontrivial process of extraction of
  - valid (with some degree of certainty)
  - novel (surprising, previously unknown)
  - potential useful
  - understandable

  patterns from large collection of data

Data Driven vs. Knowledge Driven

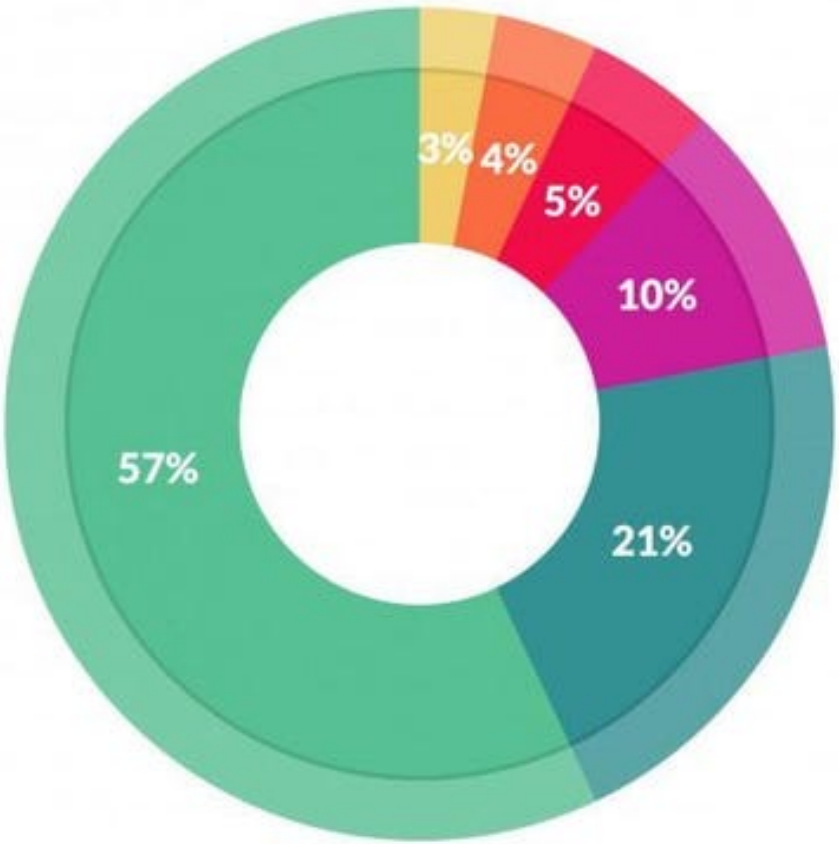
Knowledge Discovery from Data Process





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%*



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Preprocessing

- Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
  - Feature Selection (features/attributes)
  - Dimensionality Reduction (features/attributes)
  - Sampling (tuples/records/objects)

何謂資料品質好?

何謂資料品質不好?

Data Quality

- Measures for data quality
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

學生修課資料庫

Student

學號	姓名	性別	生日	系所
1101	陳綺貞	女	1975/06/06	哲學系
2301	黃奇斌	男	1990/09/10	廣告系
1102	張雨生	男	1966/06/07	外交系
1201	林志玲	女	1974/11/29	資科系
1103	謝馨儀	女	1982/04/16	企管系
1301	吳青峰	男	1982/08/32	中文系
2302	林依晨	女	1982/10/29	韓文系

SC

學生	**課程**	分數
1101	C3501	90
1102	C3001	70
1102	J2010	80
1103	C3001	100
2301	J3020	85
2301	C3001	90
2302	J2010	70
2302	C3020	80
2302	C3020	80
1301		80
1302	J2010	85

Course

代碼	名稱	學分	人數限制	選課人數
C3001	資料庫系統	3	10	2
J2010	演算法	3	5	3
C3020	人工智慧概論	3	10	2
C3501	資料庫系統	3	5	1

學生修課資料庫

inconsistency

Student

學號	姓名	性別	生日	系所
1101	陳綺貞	女	1975/06/06	哲學系
2301	黃奇斌	男	1990/09/10	廣告系
1102	張雨生	男	1966/06/07	外交系
1201	林志玲	女	1974/11/29	資科系
1103	謝馨儀	女	1982/04/16	企管系
1301	吳青峰	男	1982/08/32	中文系
2302	林依晨	女	1982/10/29	韓文系

SC

學生	課程	分數
1101	C3501	90
1102	C3001	70
1102	J2010	80
1103	C3001	100
2301	J3020	85
2301	C3001	90
2302	J2010	70
2302	C3020	80
2302	C3020	80
1301		80
1302	J2010	85

Course

代碼	名稱	學分	人數限制	選課人數
C3001	資料庫系統	3	10	2
J2010	演算法	3	5	3
C3020	人工智慧概論	3	10	2
C3501	資料庫系統	3	5	1

學生修課資料庫

Student

學號	姓名	性別	生日	系所
1101	陳綺貞	女	1975/06/06	哲學系
2301	黃奇斌	男	1990/09/10	廣告系
1102	張雨生	男	1966/06/07	外交系
1201	林志玲	女	1974/11/29	資科系
1103	謝馨儀	女	1982/04/16	企管系
1301	吳青峰	男	1982/08/32	中文系
2302	林依晨	女	1982/10/29	韓文系

SC

學生	課程	分數
1101	C3501	90
1102	C3001	70
1102	J2010	80
1103	C3001	100
2301	J3020	85
2301	C3001	90
2302	J2010	70
2302	C3020	80
1301	C3502	80
1302	J2010	85

Course

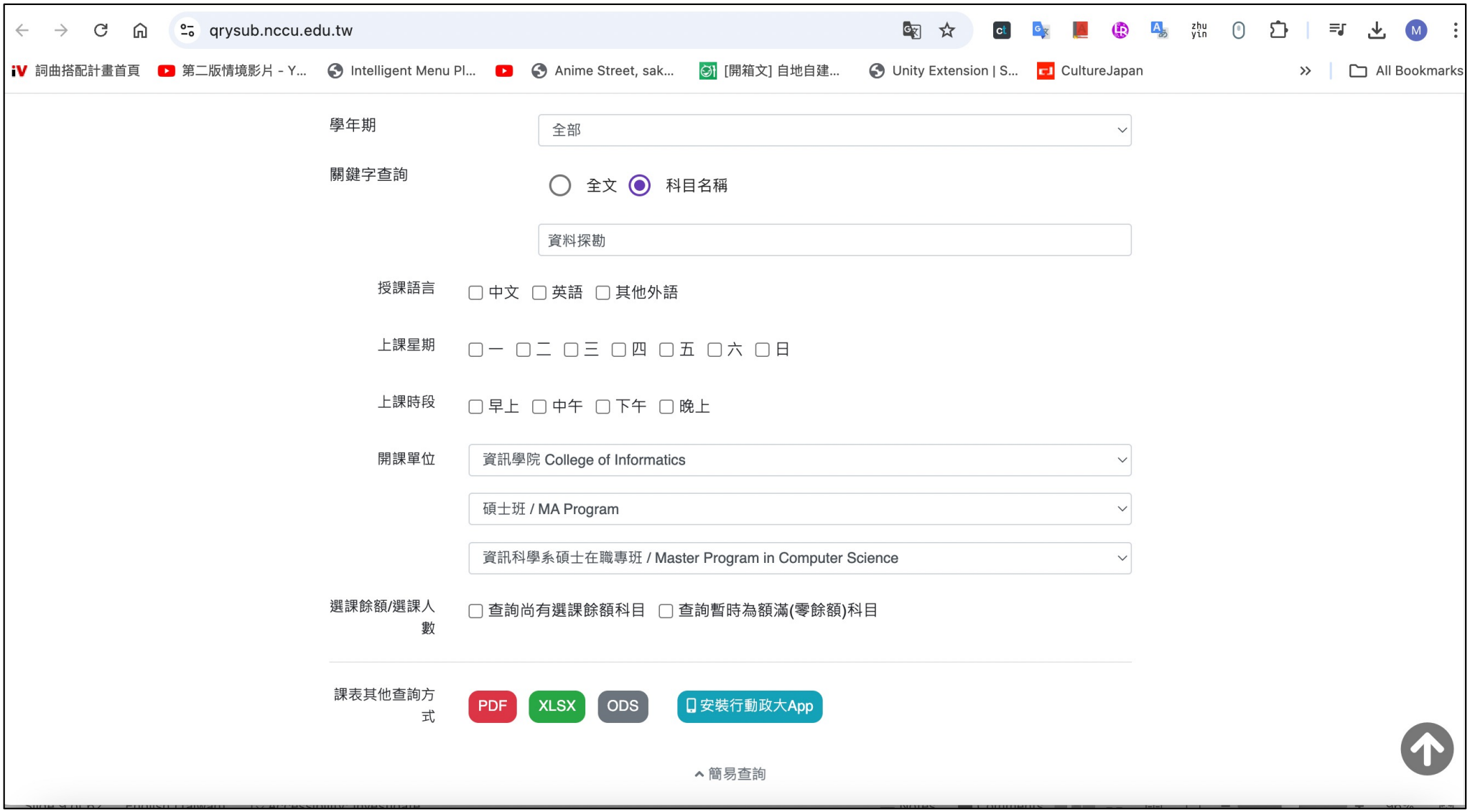
代碼	名稱	學分	人數限制	選課人數
C3001	資料庫系統	3	10	2
J2010	演算法	3	5	3
C3020	人工智慧概論	3	10	2
C3501	資料庫系統	3	5	1

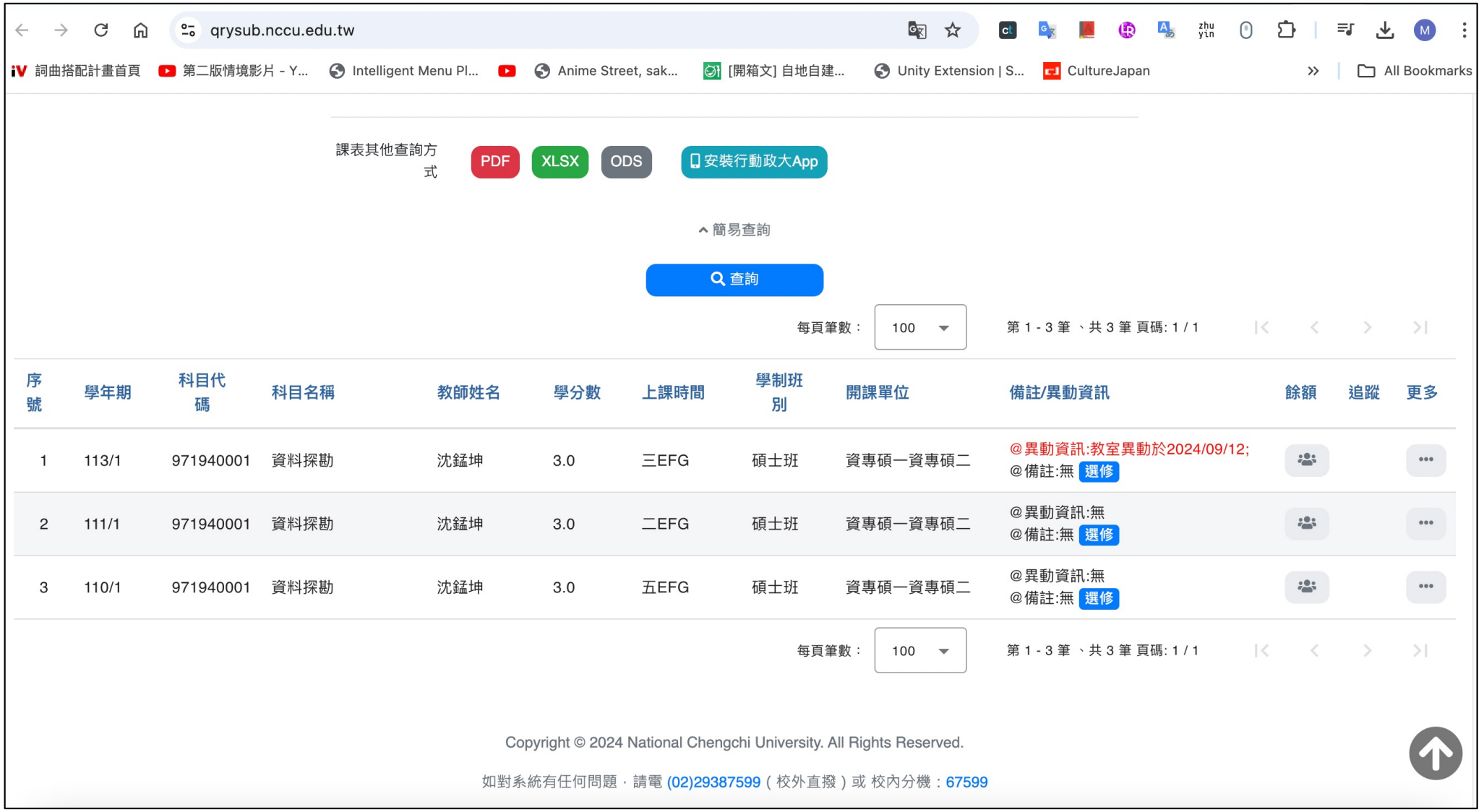
Database Management System (DBMS)

Relational Data Model

- Relational Constraints: restrictions on data that can be specified on a relational database schema
  - Domain constraints : defines valid values & data types
  - Key constraints : primary key & unique & foreign key constraints
  - Entity integrity constraints : no attribute of a PK should contain NULLs
  - Referential integrity constraints

    ↳ records cannot be inserted in the referencing relation unless they exist in the referenced relation


qrysub.nccu.edu.tw
詞曲搭配計畫頁
第二版情境影片 - Y...
Intelligent Menu Pl...
Anime Street, sak...
[開箱文] 自地自建...
Unity Extension | S...
CultureJapan
All Bookmarks
學年期
全部
關鍵字查詢
全文
科目名稱
資料探勘
授課語言
中文
英語
其他外語
上課星期
一
二
三
四
五
六
日
上課時段
早上
中午
下午
晚上
開課單位
資訊學院 College of Informatics
碩士班 / MA Program
資訊科學系碩士在職專班 / Master Program in Computer Science
選課餘額/選課人數
查詢尚有選課餘額科目
查詢暫時為額滿(零餘額)科目
課表其他查詢方式
PDF
XLSX
ODS
安裝行動政大App
簡易查詢


qrysub.nccu.edu.tw
詞曲搭配計畫首頁
第二版情境影片 - Y...
Intelligent Menu Pl...
Anime Street, sak...
[開箱文] 自地自建...
Unity Extension | S...
CultureJapan
All Bookmarks
課表其他查詢方式
PDF
XLSX
ODS
安裝行動政大App
簡易查詢
查詢
每頁筆數： 100
第 1 - 3 筆 、共 3 筆 頁碼: 1 / 1
序號 學年期 科目代碼 科目名稱 教師姓名 學分數 上課時間 學制班別 開課單位 備註/異動資訊 餘額 追蹤 更多
1 113/1 971940001 資料探勘 沈錳坤 3.0 三EFG 碩士班 資專碩一資專碩二 @異動資訊:教室異動於2024/09/12; @備註:無 選修
2 111/1 971940001 資料探勘 沈錳坤 3.0 二EFG 碩士班 資專碩一資專碩二 @異動資訊:無 @備註:無 選修
3 110/1 971940001 資料探勘 沈錳坤 3.0 五EFG 碩士班 資專碩一資專碩二 @異動資訊:無 @備註:無 選修
每頁筆數： 100
第 1 - 3 筆 、共 3 筆 頁碼: 1 / 1
Copyright © 2024 National Chengchi University. All Rights Reserved.
如對系統有任何問題，請電 (02)29387599（校外直撥）或 校內分機：67599

Sex (at Birth)

- ○ Female
- ● Male
- ○ Other
- ○ Prefer not to state

Gender Identity

- ○ Female
- ● Male
- ○ Transgender Female (MTF)
- ○ Transgender Male (FTM)
- ○ Gender Queer
- ○ Other
- ○ Prefer not to state

Sexual Orientation

- ● Heterosexual
- ○ Gay/lesbian
- ○ Bisexual
- ○ Other
- ○ Prefer not to state

Major Tasks in Data Preprocessing

- **Data Cleaning**
  - Fill in missing values
  - Smooth noisy data
  - Identify or remove outliers
  - Resolve inconsistencies
- **Data Integration**
  - Integration of multiple databases, data cubes, or files
- **Data Transformation**
  - Feature Scaling & Normalization
  - Discretization
  - Concept hierarchy generation

Data Cleaning

- Data in the Real World Is Dirty
  - Missing
    - lacking attribute values, lacking certain attributes of interest
    - e.g., Occupation = “ “
  - Noisy
    - containing noise, errors, or outliers
    - e.g., Salary = “−10”
  - Inconsistent
    - containing discrepancies in codes or names,
    - e.g., Age = 42, Birthday = “03/07/2010”
    - e.g., was rating 1, 2, 3, now rating A, B, C
    - e.g., discrepancy between duplicate records
  - Intentional
    - e.g., disguised missing data
    - Jan. 1 as everyone’s birthday?

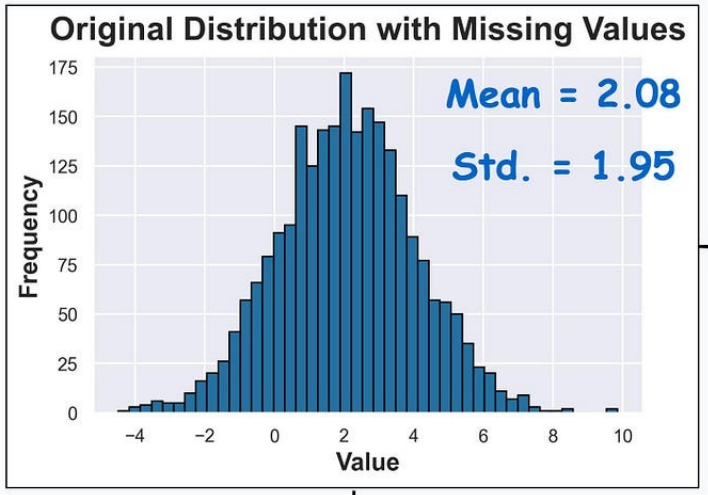
Missing Data

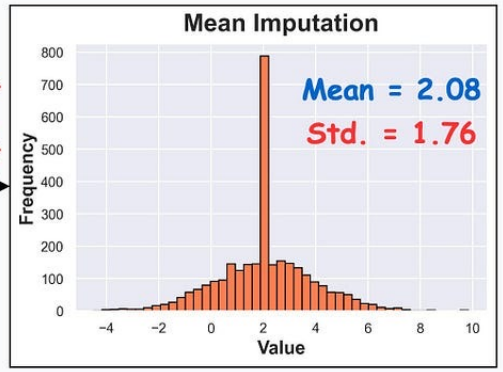
- Data is not always available
  - e.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - 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

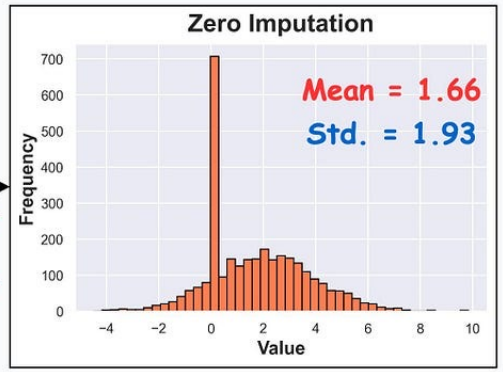
How to Handle Missing Data?

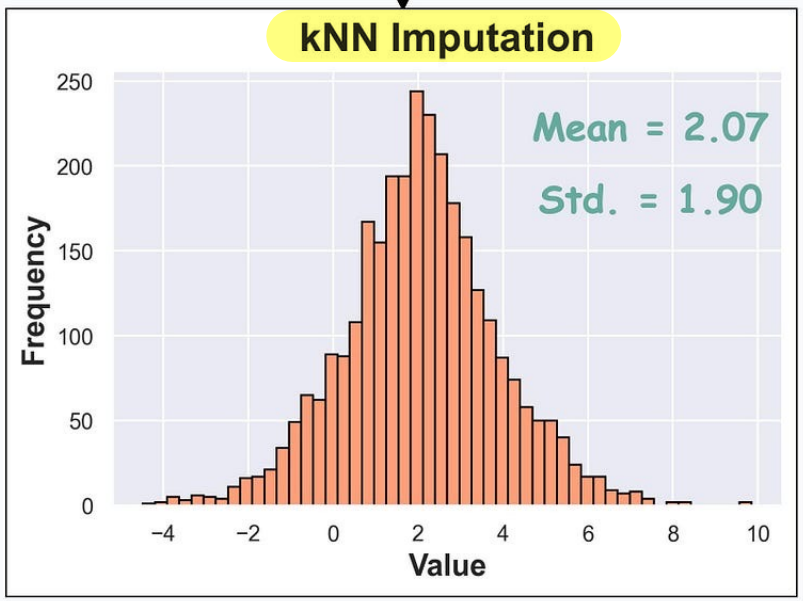
- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically (imputation) with
  - a global constant : e.g., “unknown”, a new class
  - attribute mean, median, or mode
  - attribute mean (median, mode) of similar samples, k-nearest neighbor
  - attribute mean for all samples belonging to the same class: smarter (multivariate Imputation)
  - the most probable value: inference-based by classification algorithm

Avoid Filling Missing Values With Zero or Mean



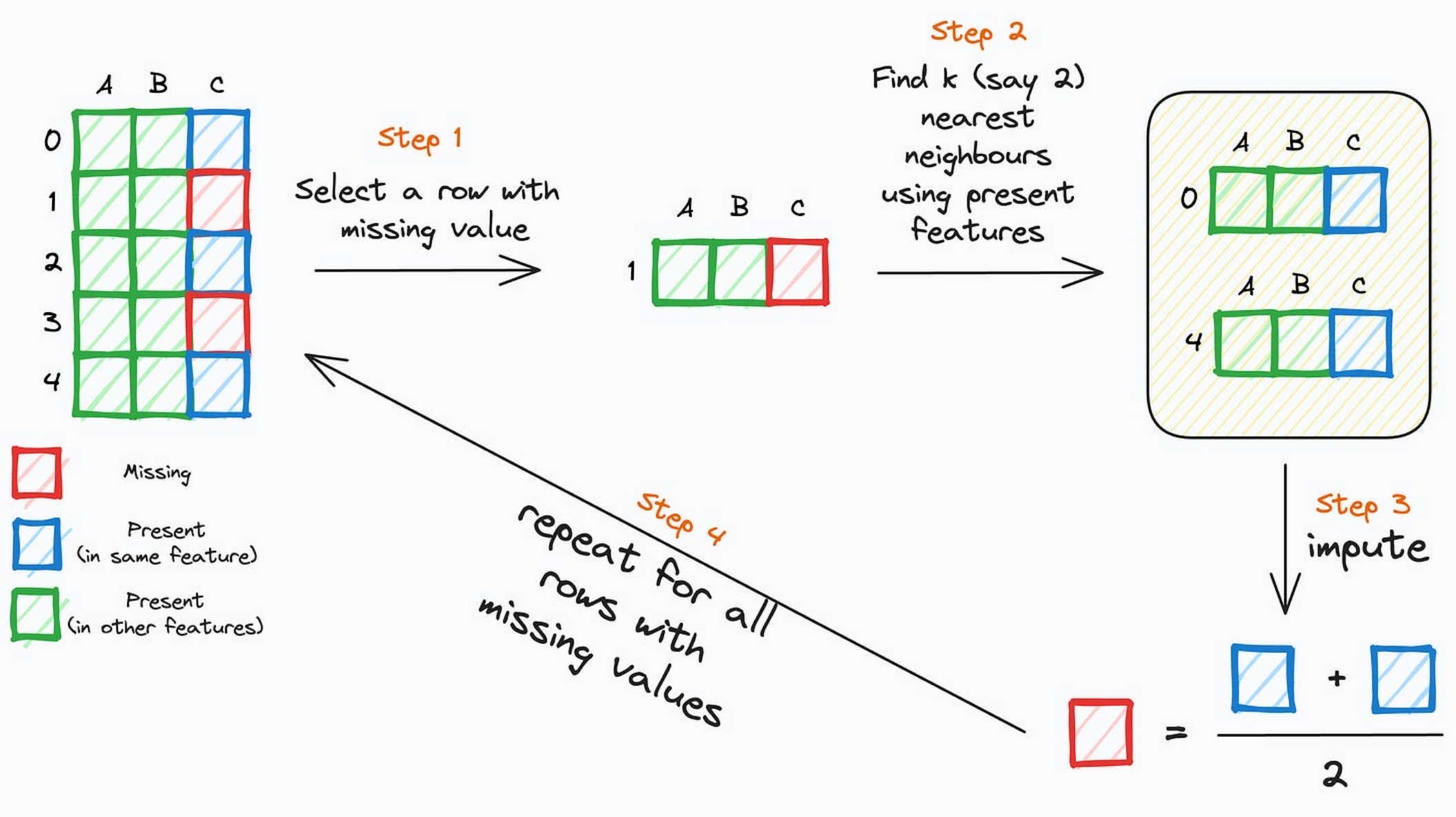






cf:https://blog.dailydoseofds.com/p/the-most-overlooked-problem-with-768

KNN Imputation



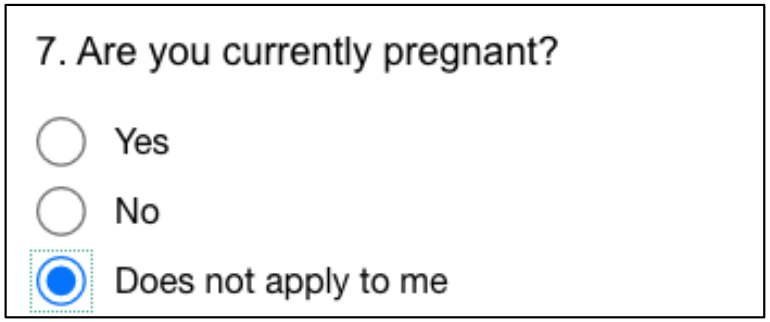
cf:https://blog.dailydoseofds.com/p/the-most-overlooked-problem-with-768

ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	**?**	60	No	1
7	Yellow	100	**?**	2
8	**?**	40	**?**	1

cf: An Evolutionary Missing Data Imputation Method for Pattern Classification

Null Values

- Null
  - Not applicable (N.A.)
    - E.g. ApartmentNumber attribute for an address of a single-family home



  - Unknown
    - Missing:
      - E.g. the Height attribute of a person
    - Not known whether the attribute value exists:
      - E.g. HomePhone attribute of a person

Null 與 0

有什麼不同？

R SATO（佐藤　玲）
@raysato

フォローする

用一張圖理解程式設計新手常搞錯的「0」跟「Null」的分別。

9,648 リツイート　7,464 いいね

4:11 - 2017年2月20日

12　9,648　7,464

葉多涵 最近人類學領域最有名的案例之一是有篇 Nature 的論文把缺值全部補成零，我同事發現如果沒這麼做的話結果會反過來，但原作者不肯認錯，期刊也不管。

28

讚 · 回覆 · 23週

王宏恩 借看

讚 · 回覆 · 23週

葉多涵 原論文：
https://www.nature.com/articles/s41586-019-1043-4
我同事的回應：https://psyarxiv.com/jwa2n/

NATURE.COM

Complex societies precede moralizing gods throughout...

6

讚 · 回覆 · 23週

Null Value (cont.)

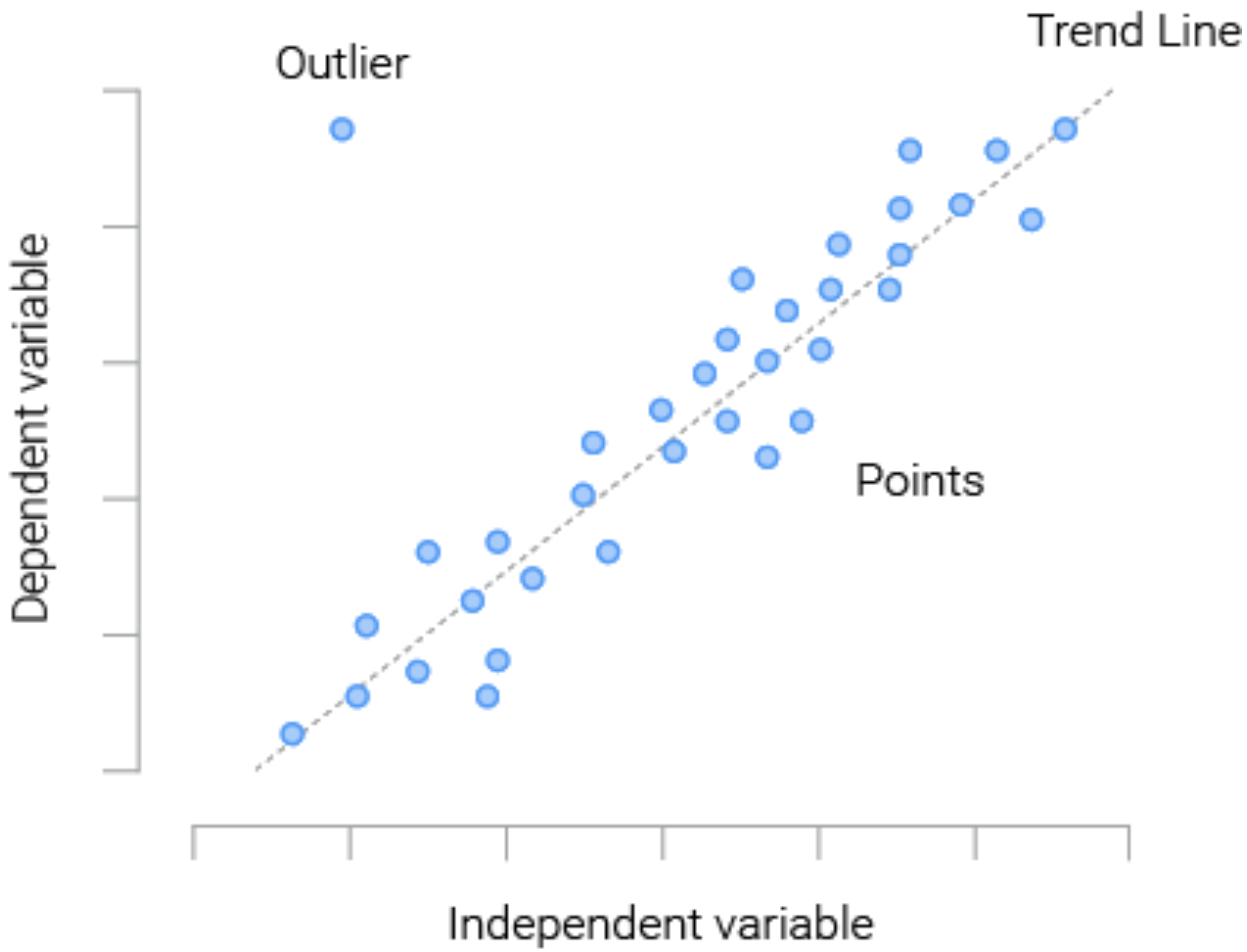
幾年前有篇跨國研究，顯示信教的小孩居然比較不願意慷慨解囊幫助別人，這篇文章獲得大量轉載。

最近有人拿這篇文章的資料重跑一次，發現作者語法寫錯了，把國家從類別變數跑成連續變數。語法更正後，結果就逆轉了，有信教的小孩稍微慷慨一點。語法錯誤的地方是STATA裡面Country應該寫成i.Country。

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

Outlier



Cf: https://ithelp.ithome.com.tw/articles/10236709

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equal-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equal-width**) bins:
  - Bin 1: 4, 8, 9
  - Bin 2: 15, 21, 21, 24
  - Bin 3: 25, 26, 28, 29, 34

$\frac{34-4}{3} = 10$

4 ... 1st boundary

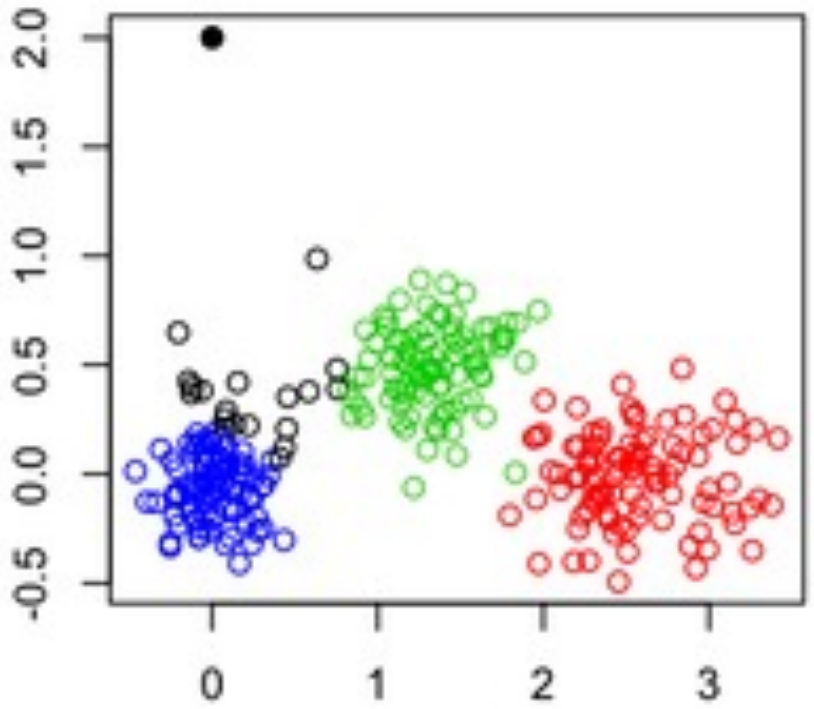
4+10=14 ... 2nd

14+10=24 ... 3rd

* Smoothing by **bin means**:
  - Bin 1: 7, 7, 7
  - Bin 2: 20.25, 20.25, 20.25, 20.25
  - Bin 3: 28.4, 28.4, 28.4, 28.4, 28.4,

* Smoothing by **bin boundaries**:
  - Bin 1: 4, 9, 9, 9
  - Bin 2: 15, 24, 24, 24
  - Bin 3: 25, 25, 25, 29, 34

Clustering & Outlier Removing



Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Integration

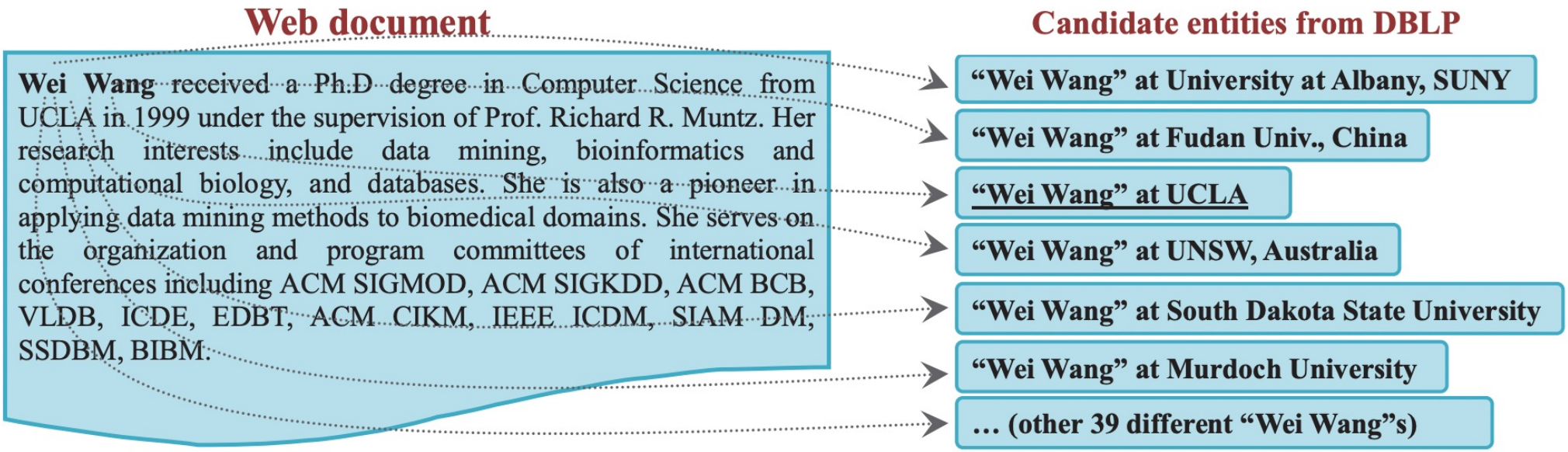
- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-# (join)
  - Integrate metadata from different sources
- Entity linking problem:
  - Identify real world entities from multiple data sources
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
- Handling redundancy

Schema Integration

- Schema Integration
  - Naming Integration
    - e.g., ID vs. Number
  - Encoding (Representation) Integration
    - e.g., 學號：理學院 vs. 資訊學院
  - Measurement Integration
    - e.g., $NT vs. $USD

Entity Linking

- Wei Wang ?



cf: A probabilistic model for linking named entities in web text with heterogeneous information networks, SIGMOD 2014.

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases
  - Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue
  - Redundant attributes: may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - e.g., Fourier transformation, wavelet transformation
  - Attribute construction
    - Combining features
    - Data discretization

Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Transformation

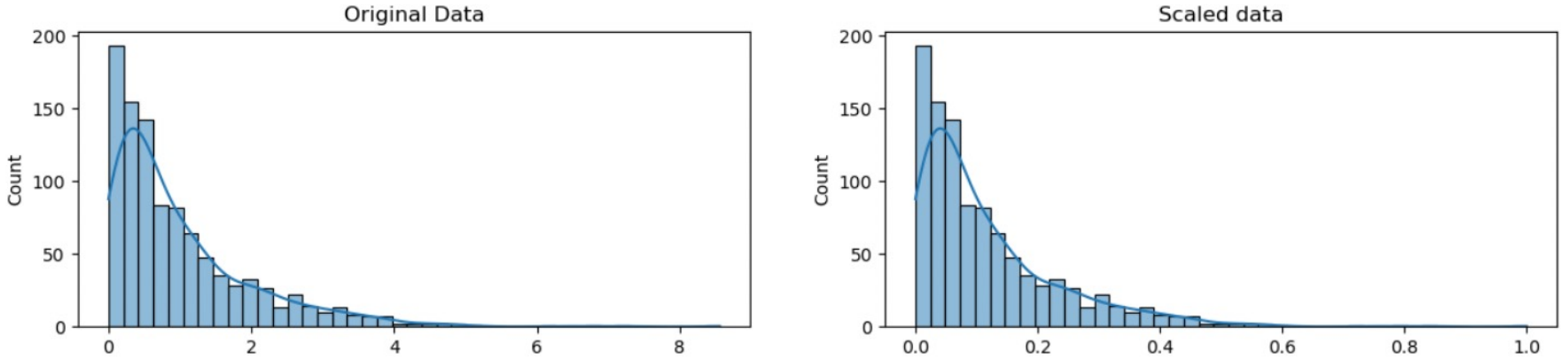
- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values

- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization
  - Feature Scaling: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

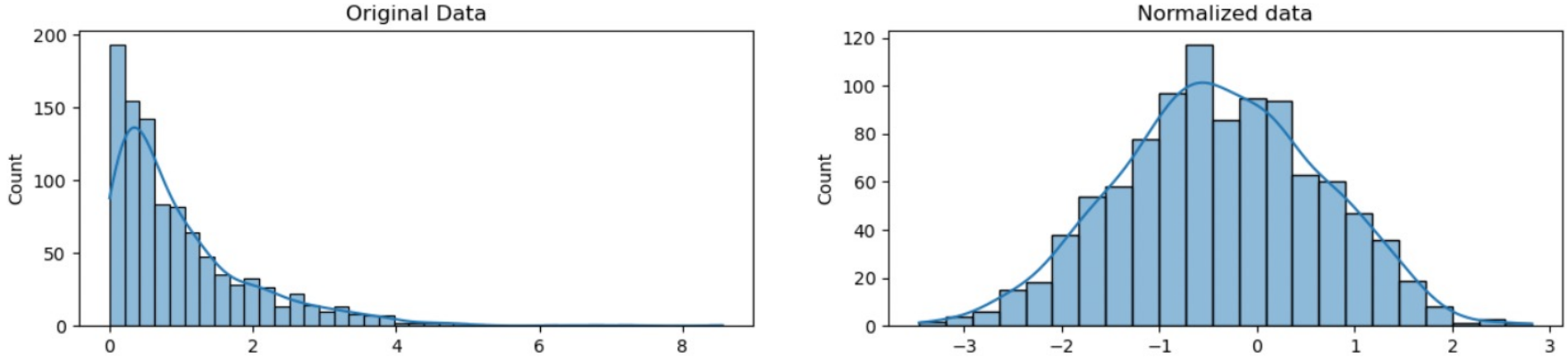
a type of discretization

Scaling vs. Normalization

- Feature Scaling: changing the range of data without changing shape of the distribution.
- Feature Normalization: changing the shape of the distribution of data



scaling



normalization

Cf: https://www.kaggle.com/code/alexisbcook/scaling-and-normalization

Feature Scaling

- **Min-max normalization** to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

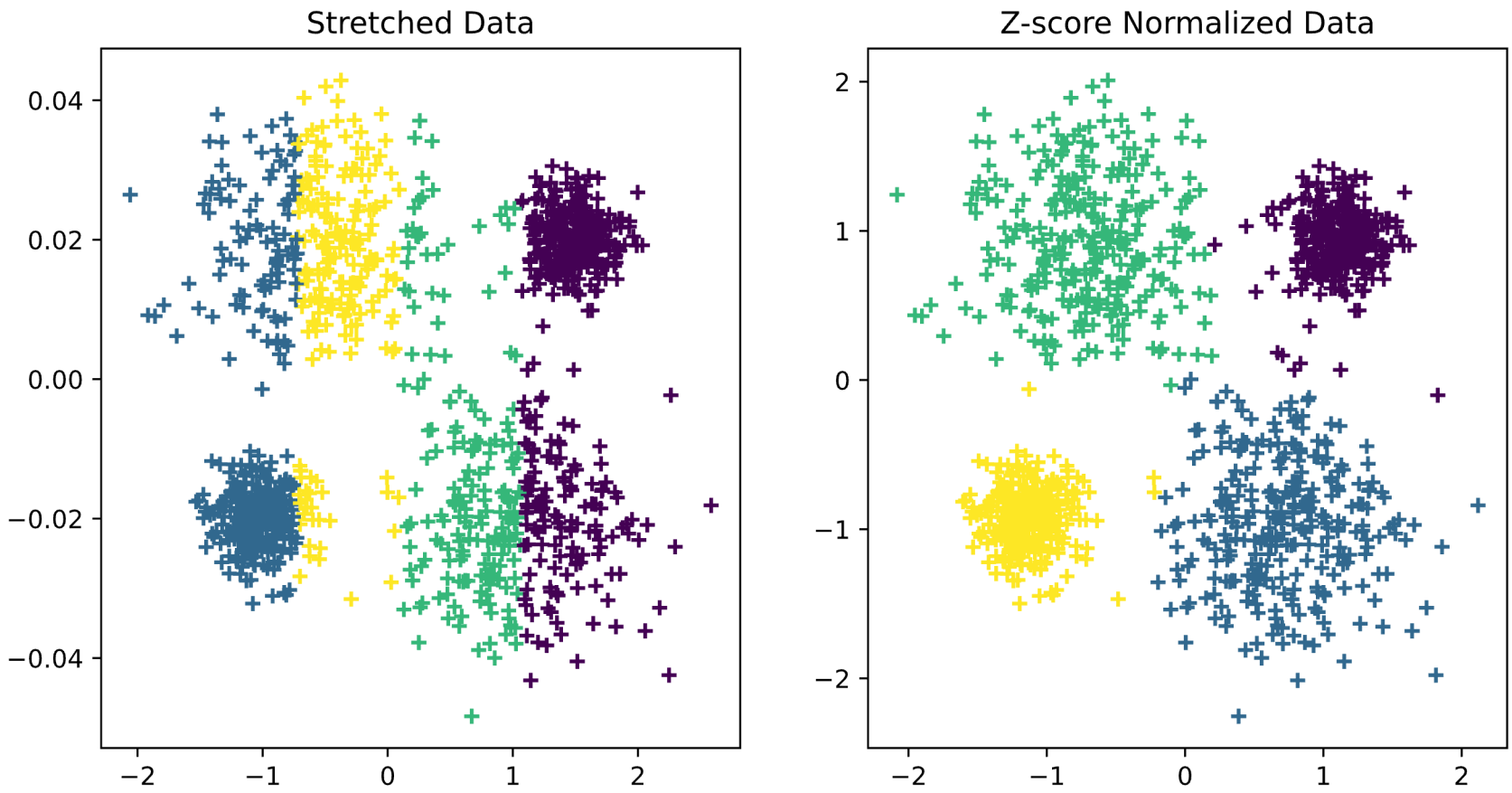
- **Z-score normalization** (μ: mean, σ: standard deviation): ≡ standardization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling** (moving the decimal point)

$$v' = \frac{v}{10^j} \quad \text{where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$



Cf: By Cosmia Nebula - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=151251915

Discretization

- Types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

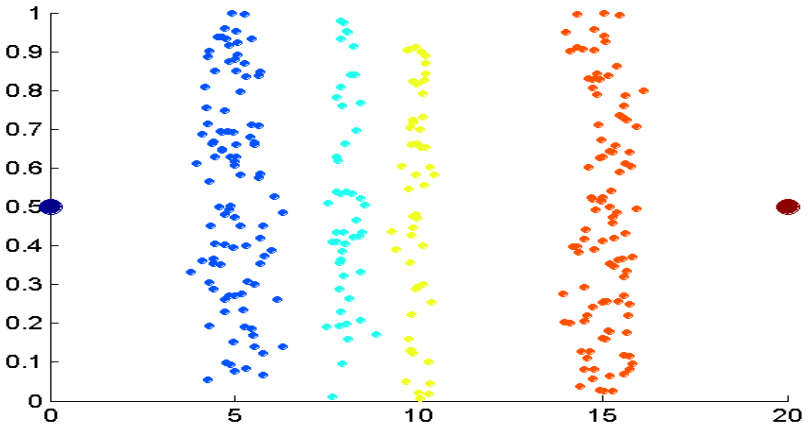
Data Discretization Methods

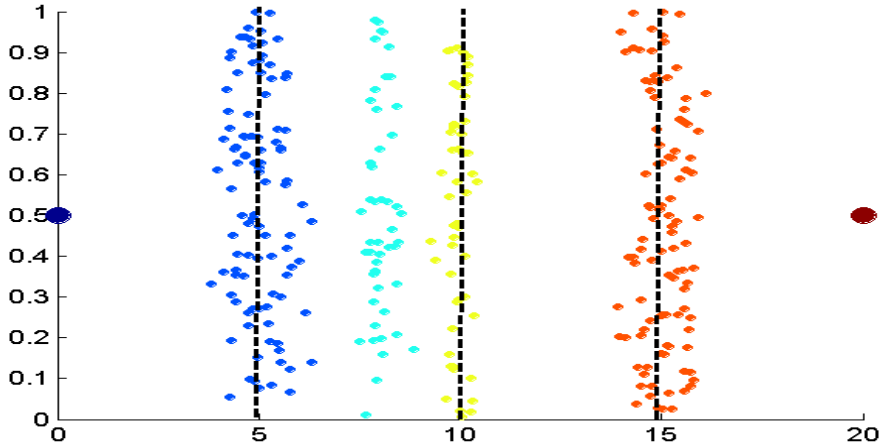
- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split
  - Histogram analysis
    - Top-down split
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)

Simple Discretization: Binning

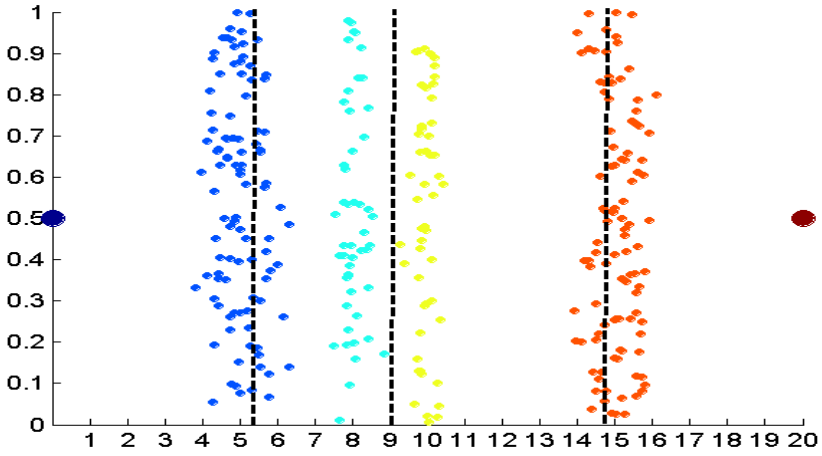
- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Discretization
Binning vs. Clustering

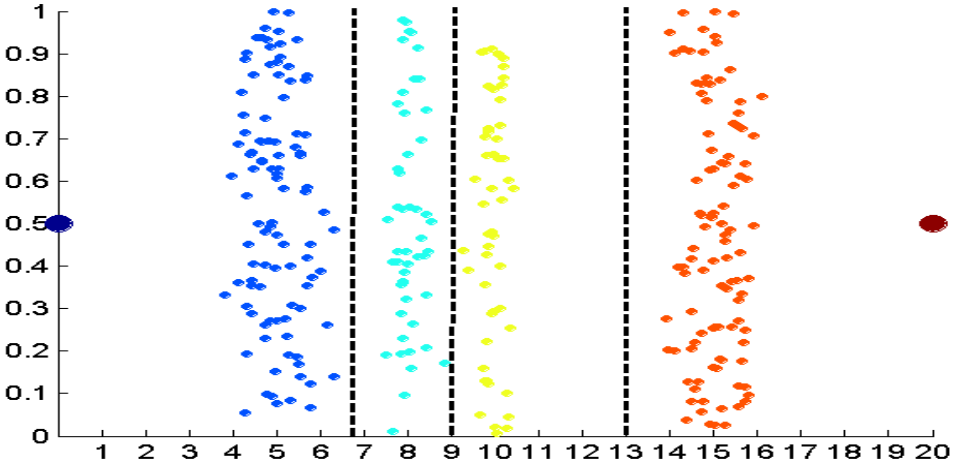




Binning: Equal Width



Binning: Equal Depth (Frequency)



K-means Clustering

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically
- e.g. department → College → University

  street → district → city → county → state

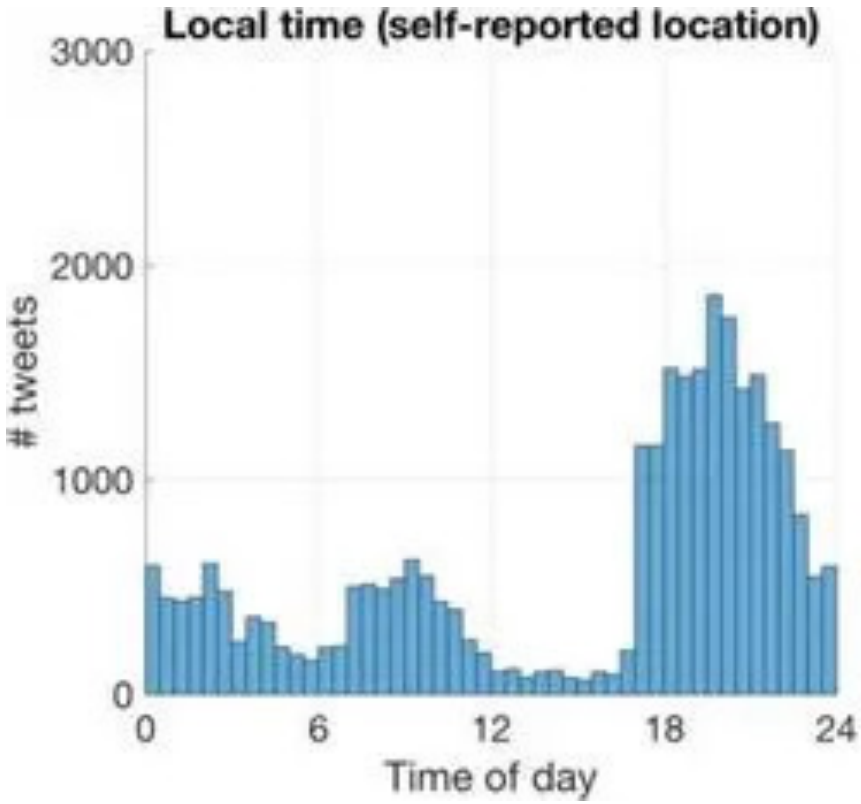
  age → {youth, adult, senior}
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods.

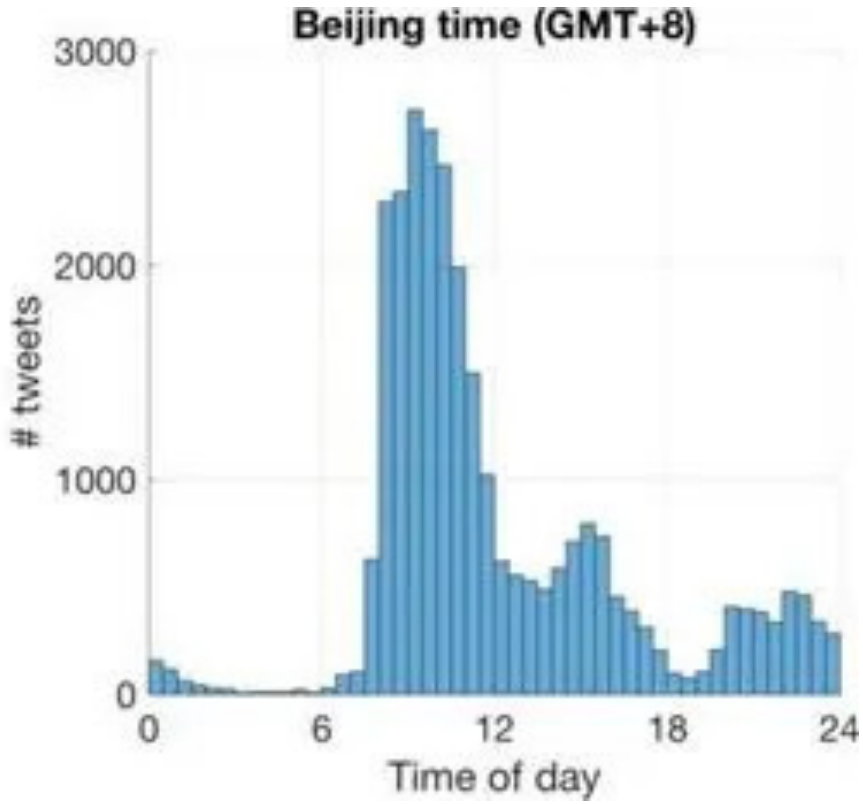
Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Date/Time Transformation

- Twitter's disclosed data on coordinated activity
  - time of day tweeted in Chinese, as local time (based on self-reported location) or Beijing time.





Cf: https://x.com/AirMovingDevice/status/1163633959441354753

Data preprocessing: Data cleaning

- Noise of ingredient: explaining

項目	解釋	範例
小括號	表示食材的狀態 (形狀/大小/數量/用途/替代品)	馬鈴薯(切小塊), 洋蔥(中型), 雞腿(共約 600 克), 醬油(雞腿肉醃料), 韓國麻油(或台式醬油), 紅蘿蔔絲(可省略)
其他括號	用於食材前表示食材的屬性 【】, [],[],], 『』, 「」, 《》, <>	<調味料>,〔材料〕
分號	常用於分割食材 ::..- _ = ~ · • , 。, * ^ ! ※	醬料:醬油 , 醃肉醬料—醬油
其他	空白 , \u200b (zero-width-space inUnicode)	韭 菜 , 洋 蔥
數字	順序, 份量, 大小 123 一二三	1.去骨雞腿肉 雞腿 1 大隻, 桔梗 2 把, 10 號雞, 16 開昆布
英文	順序, 品牌, 翻譯	a 醬油, b 蛋 costco 大長今泡菜, spam 韓式魚餅 FriedFishCakes

cf: 柳桓任: 由食譜資料探勘分析料理的在地話:以韓式料理為例

Data preprocessing: Data cleaning

- Noise of ingredient: explaining (cont.)

項目	解釋	範例
選擇	選擇 兩者選一情況 或, /, /, or	肉桂棒或桂皮 黑糖 or 紅糖
合併	將食材寫一起 和, 及, &, +	蒜末及辣椒末 青辣椒&紅辣椒
其他	料理用到器具, 生活用品 (鍋類, 機器, 小工具)	鑄鐵鍋, 湯鍋, 不沾鍋 食物調理機, 麵包機, 豆漿機 餐夾 牙籤 剪刀 膠帶 橡皮筋 清潔手套
補充說明	使用者備註	依個人喜好 糖視情況可不加

Data preprocessing: Data cleaning

- Noise of ingredient: misspelling

項目	範例
左方部首錯誤	蛤蠣 (蛤犡) 、牡蠣 (牡犡) 味噌 (味增) 、味酏 (味琳, 味啉, 味啉, 味林) 胡椒 (椥椒, 糊椒) 、杏鮑菇 (杏包菇) 、鴻喜菇 (鴻禧菇) 、青江菜 (清江菜) 、美乃滋 (美乃茲, 美奶滋)
上方部首錯誤	蘑菇 (磨菇) 、蛤蠣(蛤蟅) 、茭白筍 (筊白筍)
上方部首錯誤(草字頭)	菠菜 (波菜) 、麻油 (蔴油) 、豆芽 (豆牙, 荳芽) 、白蘿蔔 (白羅蔔)
相似字	蔥 (蔥) 、栗子(粟子) 、嫩豆腐 (嫩豆腐)
發音類似	茼蒿 (茼凹) 、在來米粉 (再來米粉) 、薑(僵)
順序顛倒	透抽 (抽透)

Data preprocessing: Synonym Processing

- Processing principles of Synonyms : based on the same entity
- 5 types of synonym
  1. Alias：地瓜 (蕃薯)，蒜 (蒜頭-大蒜)， 味酏 (味淋-味霖-米霖)
  2. Alias for Different shape：蒜 (蒜末-蒜茉-蒜泥-蒜蓉-蒜仁-蒜茸-蒜瓣-蒜粒-蒜米)
  3. Different parts of same entity：紅蔥頭(全株的頭)-珠蔥(全株)
  4. Variant Chinese character：鹽(塩)，薑(姜)，蔥(葱)， 雞(鶏)
  5. Adjective + entity：核桃 (已烤香的核桃)，山茼蒿(山茼蒿尾端葉子) ，黃豆(有機黃豆)

珠蔥

珠蔥全株

紅蔥頭

Data preprocessing: Synonym Processing

- Synonym in data

Ingredient	synonym	Ingredient	synonym
鹽	鹽巴	鴻喜菇	雪白菇, 美白菇, 本菇
蔥	青蔥, 大蔥	香菇	冬菇
櫛瓜	夏南瓜, 翠肉瓜, 翠玉瓜, 節瓜	牡蠣	蚵仔
紅蘿蔔	胡蘿蔔, 甘筍	蛤蜊	蛤蠣
地瓜	番薯	起司	起士, 芝士, 乳酪
木薯粉	樹薯粉	雞蛋	蛋
蘆筍	露筍	鵪鶉蛋	鳥蛋

13道番茄炒蛋 from 多多開伙

	烹調時間	蛋	番茄	番茄醬	糖	蒜末	薑	蔥	醬油	鹽	油	水	太白粉	洋蔥	昆布粉
1	15	3 顆	1 顆	2 大匙	少許	少許	少許		少許						
2	15	4 個	150 克	2 大匙	1 大匙			2 支			2 大匙	150cc			
3	0	5 顆	3 顆		適量		適量	2 支					適量		
4	20	4 顆	2 顆	3 大匙	1 大匙			2 支					1 大匙		
5	0	2~3 顆	2~3 顆		適量			數根	適量	適量	適量				
6	30	3 顆	1 顆	2 大匙	1 大匙			1 支		1 小匙		3 大匙	2 小匙		
7	30	3 個	1~2 顆	3 大匙	1 大匙					少許		2 大匙			
8	15	大 3 個	中 3 顆	1 大匙	1 小匙	1 小匙			2 大匙		2.5 大匙				
9	20	4 顆	牛 4 顆					2 支							
10	15	2 個	一顆					2 根	1 茶匙		少許				適量
11	15	3 顆	5		少許			少許		少許	2.5 湯匙				
12	20	4 顆	3	適量	適量					適量				適量	
13	20	V	V							V			V	V	

2017.02.03 | 新創團隊

日本最大食譜網站Cookpad收購台灣「多多開伙」搶寶島商機

日本食譜分享網站Cookpad逐漸受到亞洲觀眾注意。或許因為在台灣知名度大大提升，讓他們決定深耕寶島市場，收購台灣同類型網站「多多開伙」，納入旗下原有的台灣服務，至於多多開伙的所有會員、食譜也跟著全數轉移。

吳元熙

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**
- **Data integration**
- **Data transformation**