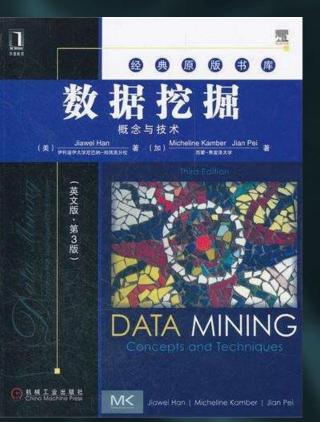
Technology Science Information Networks Computing



Lecturer: Ting Wang (王挺)

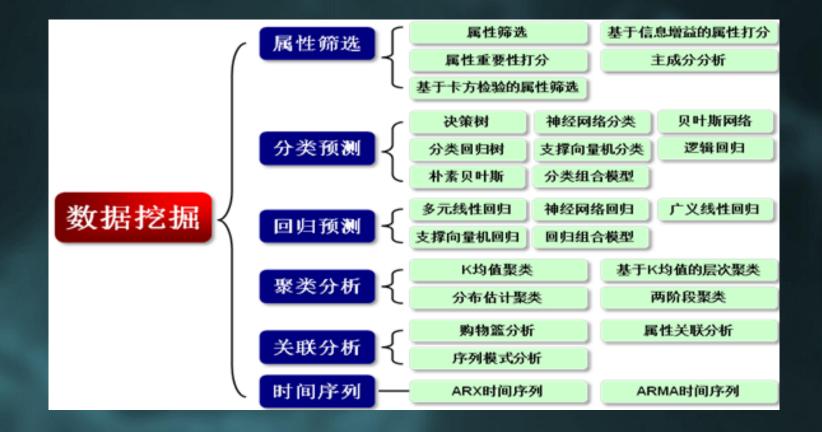
利物浦大学计算机博士 清华大学计算机博士后 电子信息技术高级工程师 上海外国语大学网络与新媒体副教授 浙江清华长三角研究院海纳认知与智能研究中心主任

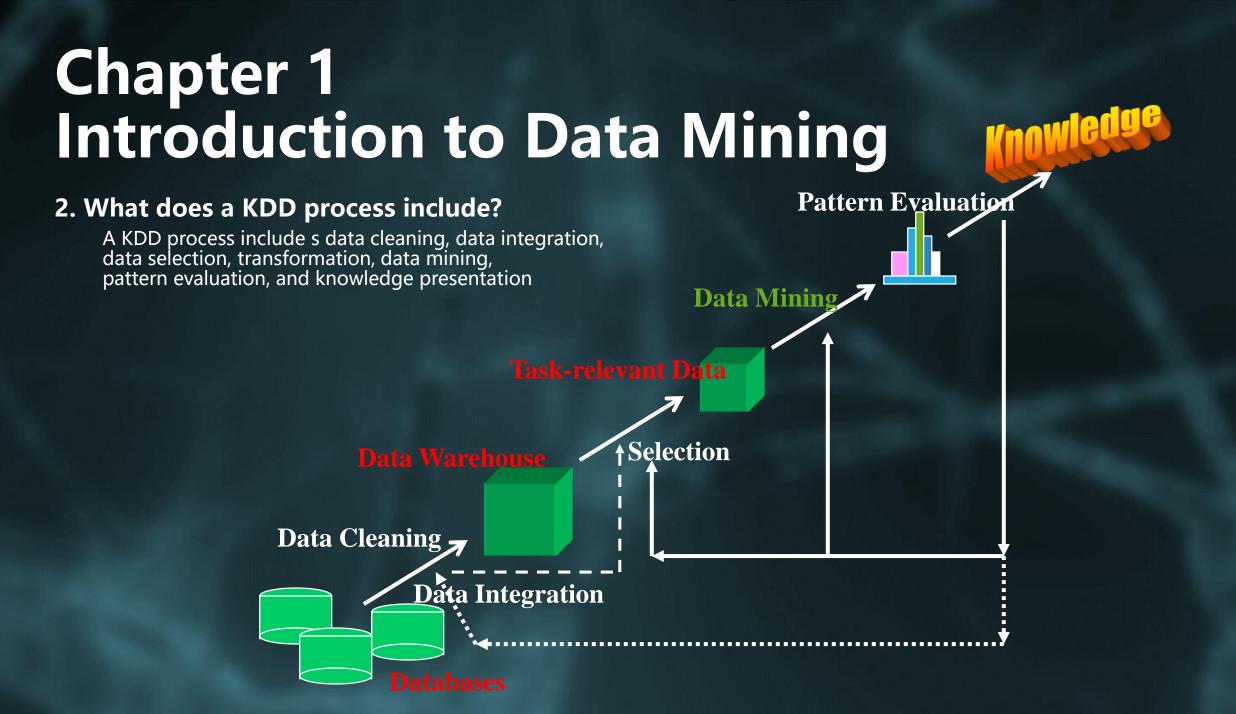




#### 1. What is Data Mining? (The Concepts of Data Mining)

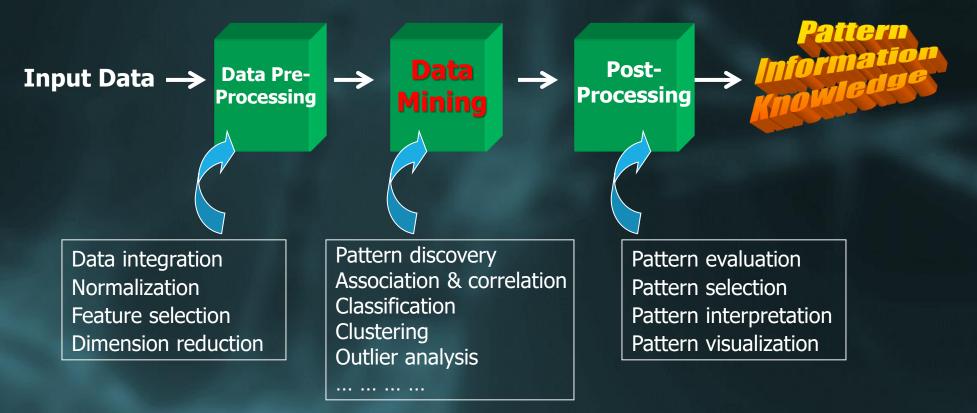
Discovering interesting patterns and knowledge from massive amount of data





### 3. What are the functionalities of data mining?

Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.



4. Do you know any application scenarios of Data Mining?
7 Best Real-Life Example of Data Mining https://prowebscraper.com/blog/data-mining-examples/



MONGOLI

GANSU

DINGHAI

JIANG

TIBET

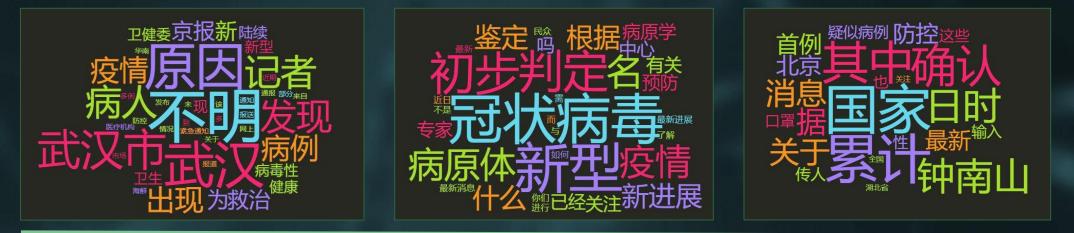
INER MONGOLIA





#### Analysis 1:

The history of Wuhan coronavirus in January 2020 http://www.myzaker.com/article/5e2e9518b15ec033014c366a/



Time

#### Analysis 2:

Five million people left Wuhan before the lockdown, where did they go? https://www.yicai.com/news/100481655.html

#### 春运期间,从武汉返乡的人群中, 回到孝感和黄冈的人群比例最高

1月10日至1月22日每天从武汉出发前往湖北省内其他城市的人群比例(单位:%)

| 孝感 - | 10.94 | 13.00 | 13.47    | 12.04    | 12.61 | 12.14 | 12.57 | 12.56 | 13.14 | 14:47 | 16.24 | 13.87 | 13.34 |
|------|-------|-------|----------|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 黄冈 - | 10.52 | 11.75 | 11.39    | 11.39    | 12.55 | 12.30 | 13.35 | 14.21 | 14.87 | 12.28 | 12.45 | 13.50 | 12.95 |
| 荆州-  | 5.74  | 5.91  | 5.74     | 5.80     | 5.84  | 6.03  | 6.00  | 5.93  | 6.29  | 6.93  | 7.29  | 7.17  | 7.80  |
| 咸宁-  | 5.22  | 5.95  | 5.32     | 4.94     | 4.97  | 5.10  | 4.96  | 5.07  | 5.14  | 4.95  | 4.75  | 4.77  | 4.38  |
| 鄂州-  | 4.12  | 4.53  | 4.83     | 4.77     | 4.36  | 4.10  | 4.04  | 4.23  | 4.39  | 3.91  | 3.53  | 3.28  | 3.26  |
| 襄阳-  | 4.12  | 3.92  | 3.66     | 3.72     | 3.68  | 3.44  | 3.44  | 3.58  | 3.63  | 3.81  | 4.08  | 4.44  | 4.74  |
| 黄石一  | 3.42  | 3.81  | 3.74     | 3.70     | 3.69  | 3.68  | 3.84  | 3.94  | 4.21  | 3.75  | 3.70  | 3.74  | 3.40  |
| 荆门一  | 2.85  | 2.95  | 2.72     | 2.76     | 2.73  | 2.82  | 2.81  | 2.75  | 2.96  | 3.31  | 3.59  | 3.76  | 3.91  |
| 随州-  | 2.52  | 2.71  | 2.65     | 2.67     | 2.68  | 2.82  | 2.89  | 2.98  | 3.11  | 3.21  | 3.38  | 3.54  | 3.66  |
| 仙桃 - | 2.38  | 2.81  | 2.80     | 2.66     | 2.59  | 2.88  | 2.80  | 2.76  | 2.91  | 3.07  | 3.11  | 3.23  | 3.19  |
| 宜昌-  | 3.08  | 3.24  | 2.76     | 2.76     | 2.43  | 2.35  | 2.48  | 2.50  | 2.54  | 2.69  | 2.95  | 3.05  | 3.49  |
| 天门   | 1.47  | 1.76  | 2.01     | 1.77     | 1.95  | 1.97  | 2.07  | 1.95  | 2.10  | 2.33  | 2.43  | 2.28  | 2.28  |
| 恩施 - | 2.12  | 1.92  | 2.11     | 1.83     | 1.89  | 1.79  | 1.80  | 1.74  | 1.80  | 1.74  | 1.87  | 1.83  | 1.80  |
| 十堰-  | 2.02  | 1.85  | 1.88     | 1.76     | 1.65  | 1.60  | 1.50  | 1.56  | 1.65  | 1.84  | 1.97  | 2.00  | 1.99  |
| 潜江-  | 1.10  | 1.14  | 1.28     | 1.18     | 1.04  | 1.01  | 1.03  | 1.02  | 1.03  | 1.04  | 1.17  | 1.19  | 1.43  |
| L_   |       | ····  | ····r··· | ····r··· | ····r | r     | r     | r     | r     | ····r | ····r | r     | ····  |
|      | 1.10  | 1.11  | 1.12     | 1.13     | 1.14  | 1.15  | 1.16  | 1.17  | 1.18  | 1.19  | 1.20  | 1.21  | 1.22  |



数据来源:百度地图慧眼百度迁徙 注:统计时间为2020年1月10日至2020年1月22日

| 地区    | 确诊    | 死亡   | 治愈   |
|-------|-------|------|------|
| ▼ 湖北  | 33366 | 1068 | 2646 |
| 武汉    | 19558 | 820  | 1379 |
| 孝感    | 2751  | 45   | 179  |
| 黄冈    | 2398  | 54   | 338  |
| 随州    | 1129  | 12   | 56   |
| 荆州    | 1110  | 21   | 80   |
| 襄阳    | 1088  | 12   | 60   |
| 黄石    | 874   | 6    | 96   |
| 鄂小    | 861   | 28   | 72   |
| 宜昌    | 784   | 8    | 62   |
| 荆门    | 696   | 24   | 81   |
| 十堰    | 536   | 1    | 70   |
| 咸宁    | 525   | 6    | 65   |
| 仙桃    | 460   | 13   | 43   |
| 天门    | 293   | 10   | 12   |
| 恩施州   | 203   | 3    | 38   |
| 潜江    | 90    | 5    | 8    |
| 神农架林区 | 10    |      | 7    |
|       |       |      |      |

2020年2月12日湖北疫情 https://ncov.dxy.cn/ncovh5/view/pneumonia More people went to Xiaogan and Huanggang from Wuhan, which indicates that more people in these two cities would get virus. Reports on February 12 verified this.

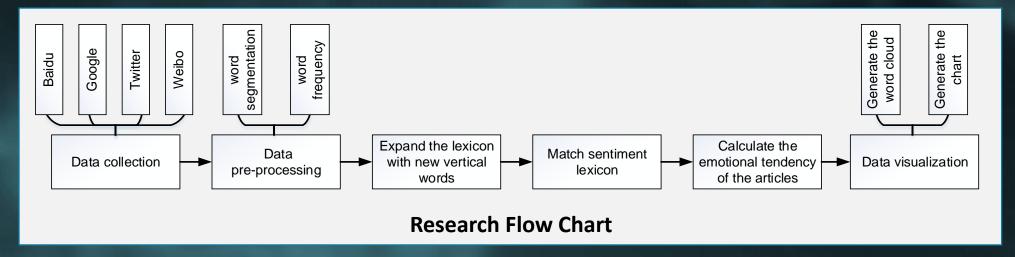


### 深圳科学家贺建奎宣布世界首例 免疫艾滋病的基因编辑婴儿在中国诞生

On November 26, 2018, the so-called "anti-AIDS" gene editing baby was born in China, which has aroused wide discussion and controversy internationally.

This study takes gene editing baby event as the main research topic, investigates the emotional tendency of Chinese and Western people towards this issue.

- This research was done by Miss Xinyu LIU and Dr. Ting WANG in 2019.
- Introduction to the Authors:
  - I. Miss Xinyu LIU is now an undergraduate student in Shanghai International Studies University (SISU), and will be a postgraduate student in Journalism and Communication at SISU in September 2020.
  - II. Dr. Ting WANG is the supervisor of Miss Xinyu LIU.



#### **Data Description**

Data

Collection

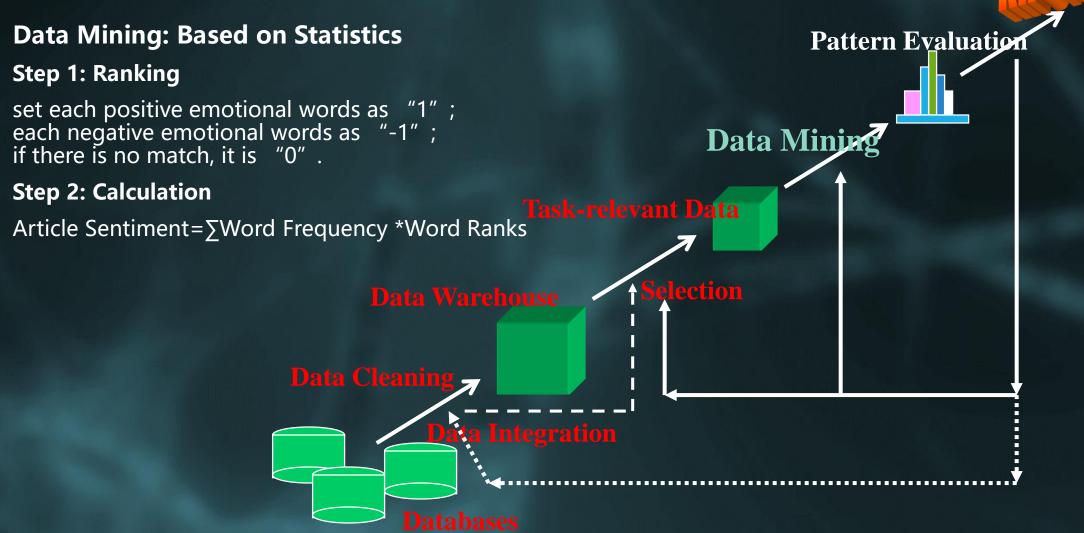
Baidu: 740

Google: 131 Weibo: 500 Twitter: 131

the study uses the web crawlers to web information from **26 November, 2018 to 15 December, 2018**, with the keyword **"gene editing baby** ". **duplicated information** has been discarded after collection.

Data Mining **Sentiment Matching** HowNet **Task-relevant Data** 18,175 emotional words Data 9,013 positive 9,162 negative Preprocessing Word quantity **†**Selection **Data Warehouse** Baidu: 195,710 Google: 44,080 Weibo: 24,755 Twitter: 3,409 Data Cleaning Data Integration Databases

Pattern Evaluation



#### **Pattern Evaluation**

|         | Negative | Neutral | Positive |
|---------|----------|---------|----------|
| Baidu   | 55.14%   | 42.7%   | 2.03%    |
| Google  | 9.92%    | 38.93%  | 51.15%   |
| Weibo   | 69.20%   | 22.60%  | 8.20%    |
| Twitter | 10.69%   | 41.22%  | 48.09%   |

Pattern Evaluation

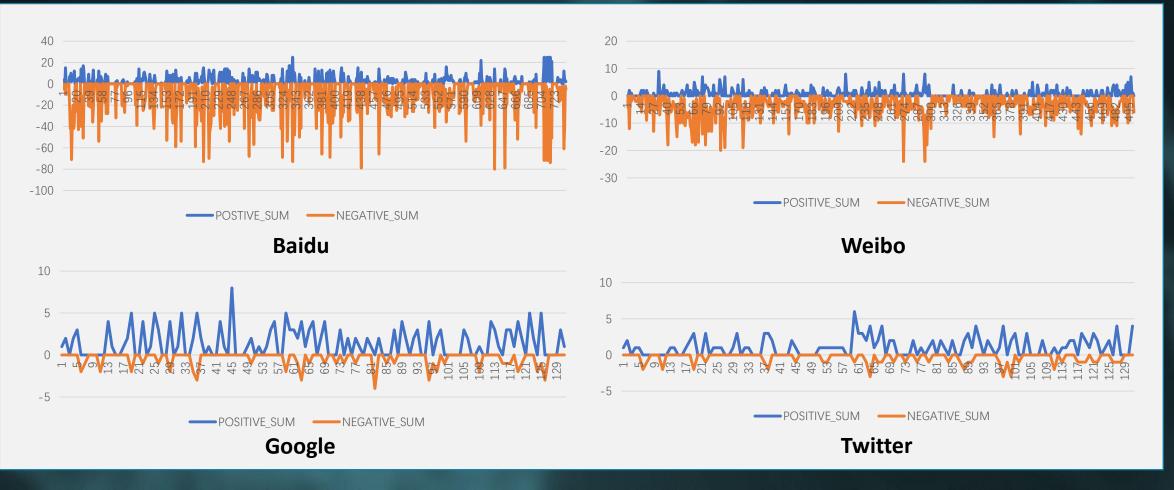
**Data Mining** 

Data Warehouse

**Task-relevant Data** 

Data Cleaning Data Integration

#### **Pattern Evaluation**





Google



Twitter

首例免疫情 研究工作制 他们的意思 婴儿中国 <sup>耕</sup> Cas9 基因 田 在 Case 技术 医学 伦理 编辑 的 进行 预建奎 副教授 胚胎 进行 Cas9 基因 基因组 编辑 健康诞生 <sup>核改 要因</sup> 学术 委员会要回 ccrs 婴儿 出生 医院 伦理 婴儿事件 #日中国 婴儿事件 专家 委员会 漸多边風感 编辑人类 中国健康保 诞生 双胞胎 生物医学研究出生 (大卫生健康 深圳市 医 生物医学研究 出生 (大加) 医激 地们出生" 这项研究 伦理 委员会伦理学术 中国 避 知情同意书 南方科技研究项目 堅決反対 個指导原则 經認效应 基因经过 節 相关信息疾病预防 深圳 美妇料 開業 III dir 基因手术離離 Numere 雪市 派 於后十年期題 首例 基因 "中 町市寺" CRISPR Cas9 對 世界 首例 通知 遭责<sup>展期</sup>所 双胞胎一个 5 kg 目前 进行 基因 <sup>10</sup> 市 学术 伦理 電源 5 kg 严重 违等 基因 测序 is " 重 抵抗 艾滋病 副 5 kg 值 // 編集 人类 基因组和 <sup>10</sup> g // 电 10 kg + 10

Baidu

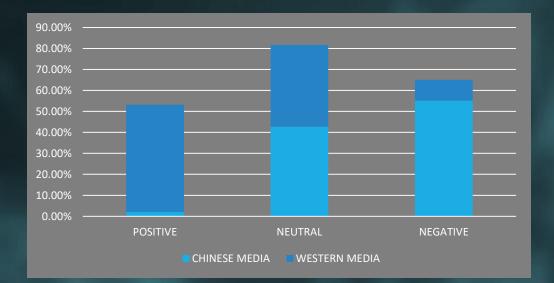


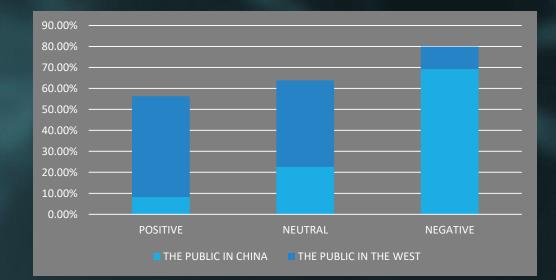
#### Pattern Evaluation

Weibo

### **Conclusions:**

- Chinese and western people have great differences in their emotional attitudes towards gene edited baby event.
- Western media and the public mainly talk about the technology and development related to gene editing.
- Chinese pays more attention to the social impacts of the incident and considers it from the perspective of ethics.





# Next>>Chapter 2

www.wangting.ac.cn