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

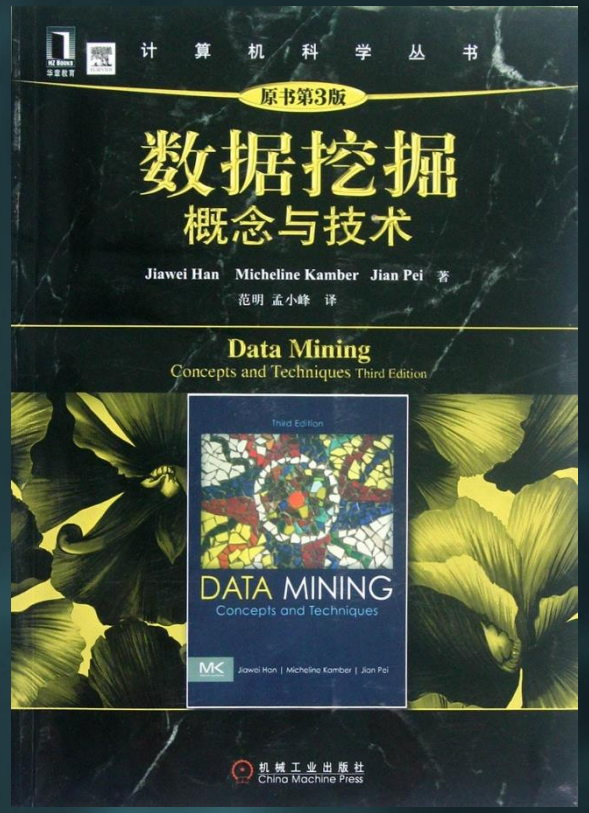
电子信息技术高级工程师

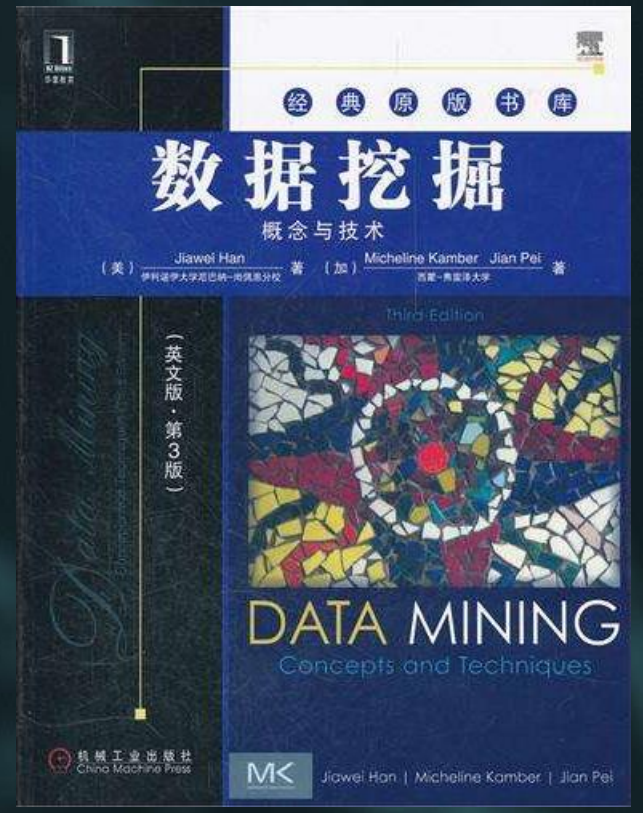
上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任

Chapter 3

Data Preprocessing





Chapter 3: Data Preprocessing

1. The purpose of Data Preprocessing: data quality

2. Data quality:

- Accuracy（精确性）: correct or wrong, accurate or not
- Completeness（完整性）: not recorded, unavailable, …
- Consistency（一致性）: some modified but some not, dangling, …
- Timeliness（时效性）: timely update?
- Believability（可信度）: how trustable the data are correct?
- Interpretability（可解释性）: how easily the data can be understood?

Chapter 3: Data Preprocessing

3. Data cleaning: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

4. Data integration: Integration of multiple databases, data cubes, or files

5. X2 (chi-square) test 卡方检验

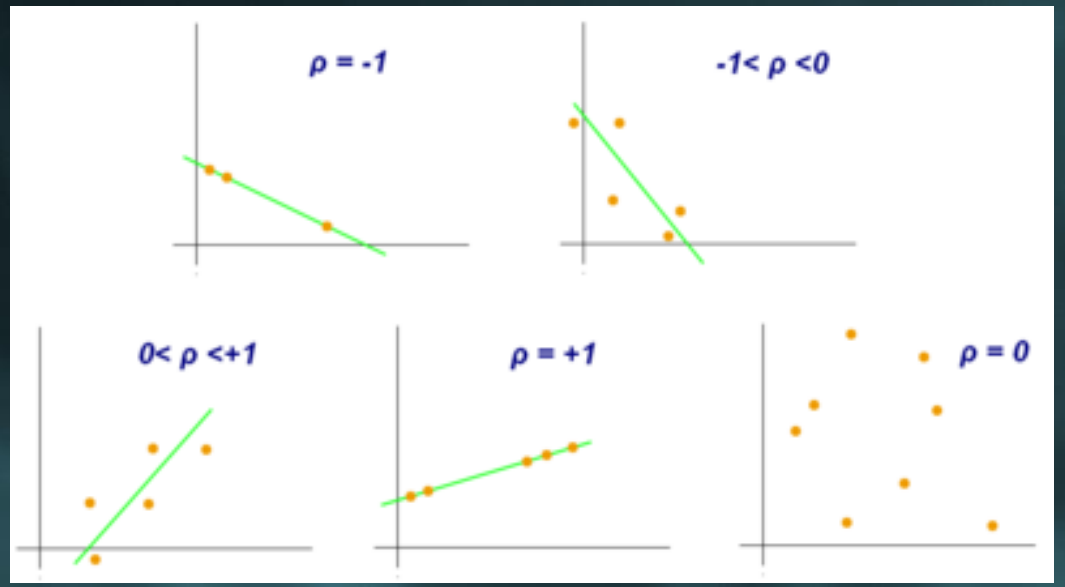
$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

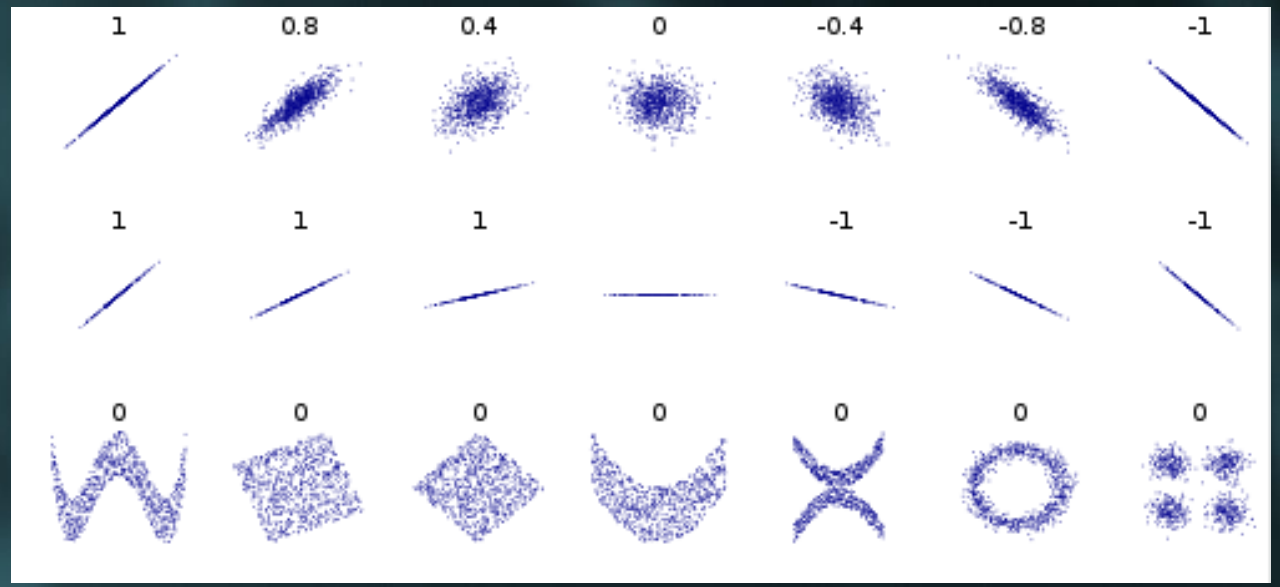
Chapter 3: Data Preprocessing

6. Pearson correlation coefficient

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y},$$

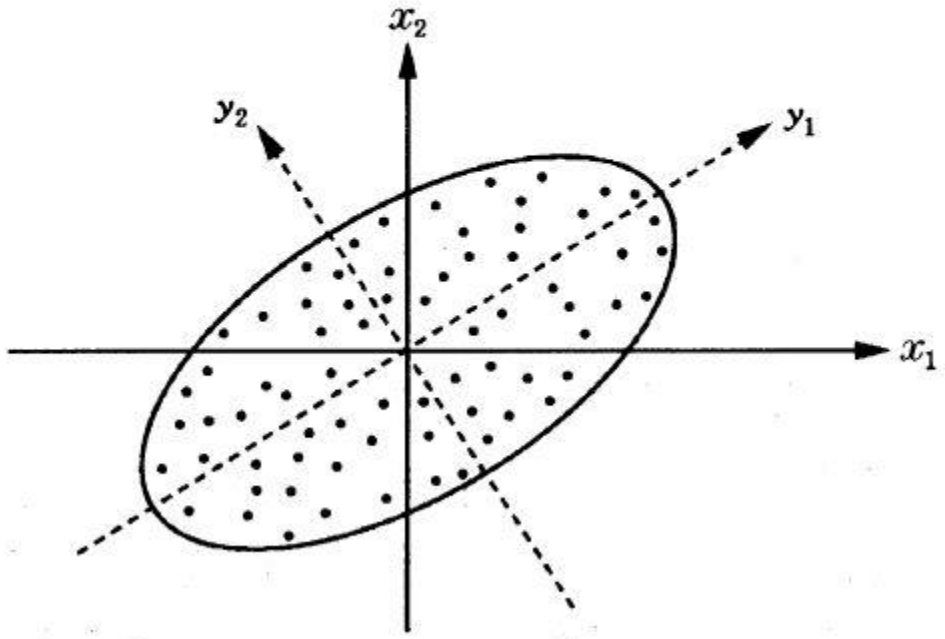


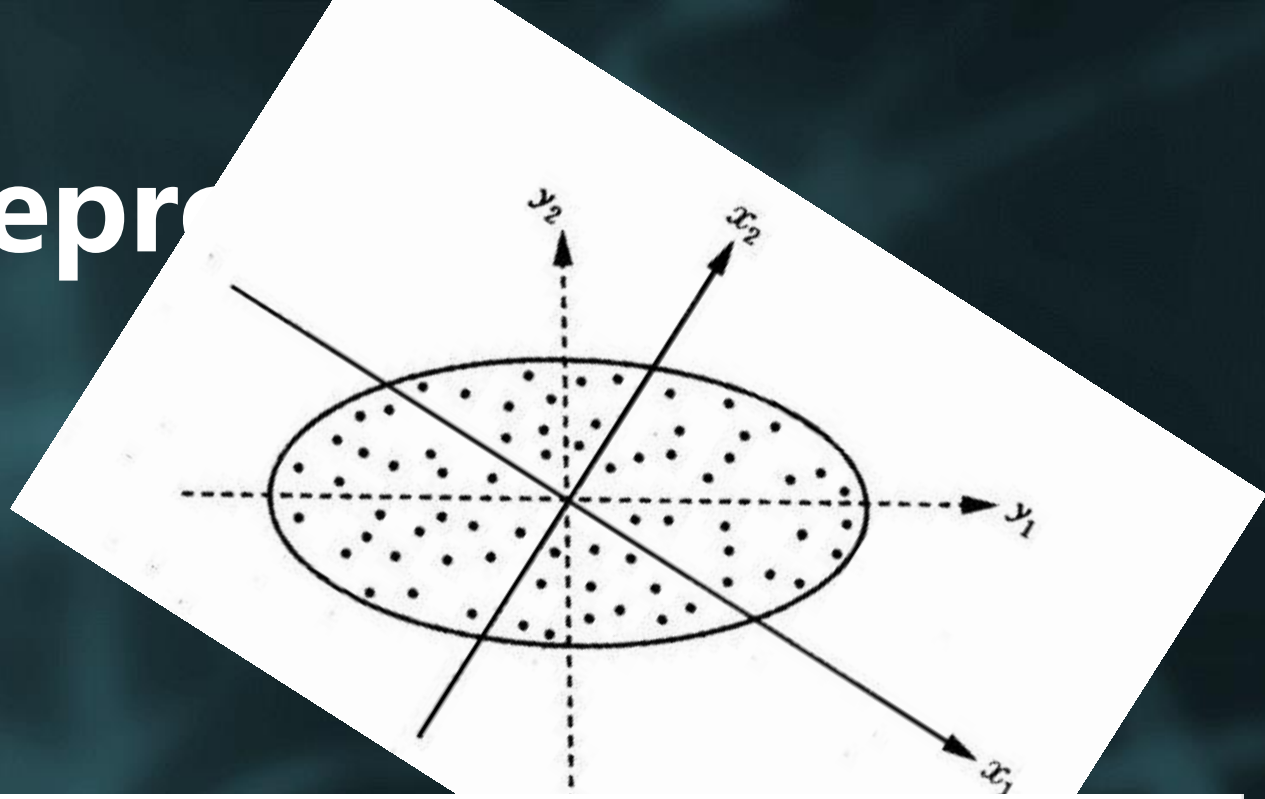


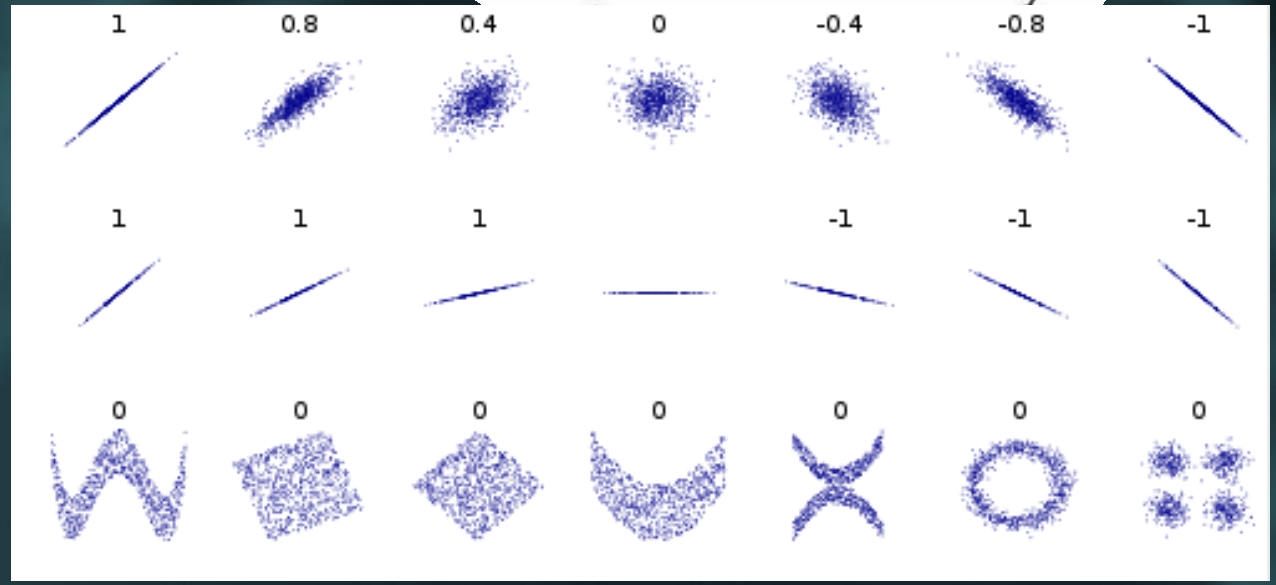
Chapter 3: Data Prepr

7. PCA(主成分分析)

$$\begin{cases} y_1 = x_1\cos\theta + x_2\sin\theta \\ y_2 = -x_1\sin\theta + x_2\cos\theta \end{cases}$$







Chapter 3: Data Preprocessing

8. Types of Sampling

- **Simple random sampling （简单随机）**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement （无放回）**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement（有放回）**
  - A selected object is not removed from the population
- **Stratified sampling（分层抽样）**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used for skewed data

Chapter 3: Data Preprocessing

9. Normalization

- Min-max normalization
- Z-score normalization
- Normalization by decimal scaling

10. Discretization

- Binning 装箱
  - Top-down split, unsupervised
- Histogram analysis 柱状图分析
  - Top-down split, unsupervised
- Clustering analysis (unsupervised, top-down split or bottom-up merge)
- Decision-tree analysis (supervised, top-down split)
- Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Chapter 3: Data Preprocessing

Chapter 3: Data Preprocessing

Data Description:

某化妆品牌微信公众号数据

2016-2017年，共计110条。

Objectives:

分析影响文章影响力的因素

a) 哪些标签会影响文章
b) 哪些代言人值得雇佣
c) 哪些产品值得去推广
d) 哪些人群应该去关注

Chapter 3: Data Preprocessing

位置	代码	星期	阅读量	转发	收藏	点赞	留言	新关注人数	取消关注人数	累计关注人数	节日	节气	热门影视	品牌活动
1	2-1	3	347	6	1	11	3	5	45	26228	1	0	1	0
1	2-2	5	592	51	0	7	99	6	41	26153	1	0	0	0

中草药	东方文化	热门话题	中奖公布	品牌价值	护肤知识	生活方式	心灵鸡汤	时尚穿搭	跨界合作	品牌促销	测评	美肌课堂	野果	红景天
0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
0	0	1	0	0	1	0	0	0	0	0	0	0	1	1

山茶花	百合	芯净	男士	四倍蚕丝	睡莲	金银花	四倍多萃	橙花	核桃	芍药	莲花	红石榴	黑茶	紫芝
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

金缕梅	防晒	光彩立现	佟丽娅	刘诗诗	王大陆	刘烨	留言评论	点赞最多	淘宝优惠	随机抽取	淘宝福利	问答	征集	品牌H5
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Chapter 3: Data Preprocessing

词云

Chapter 3: Data Preprocessing

TAG由人工预先标注

TAG主要包括:

- 时间
- 主要内容标签
- 各类活动标签
- 代言人
- 品牌与产品子类

	阅读量	转发	收藏	点赞	留言	新关注人数	取消关注人数	累计关注人数
周1	0.269502	0.3357	0.039024	0.340513	0.136431	0.068228	-0.066544	0.08068
周2	-0.042677	-0.049262	0.192446	-0.047547	-0.155564	-0.06904	-0.022823	-0.118803
周3	0.244848	0.052897	-0.063814	0.060644	0.165731	-0.076259	-0.148839	0.003213
周4	-0.146418	-0.109343	-0.01217	-0.140286	-0.15774	-0.174254	0.107741	0.033453
周5	-0.144491	-0.069997	-0.060795	-0.101508	0.210448	0.296693	-0.063813	-0.011794
周6	-0.05639	-0.075024	-0.062644	0.014647	-0.171514	-0.02628	0.129675	0.072132
周7	-0.082199	-0.057213	-0.022108	-0.055807	-0.165303	-0.106638	0.14994	-0.06983
节日	-0.231834	-0.194147	0.095221	-0.140434	-0.087887	0.022649	0.154918	0.448004
热门影视	0.206712	-0.019433	0.272237	0.007611	-0.042828	-0.163557	-0.088363	0.059642
品牌活动	-0.038851	0.067891	-0.121153	0.015275	-0.318874	0.107038	0.290171	0.081486
留言评论	-0.027292	0.010425	-0.032981	-0.076751	0.176103	-0.057623	0.14896	0.232518
征集	-0.075834	-0.050919	-0.021341	-0.09243	-0.085898	-0.030542	-0.054888	0.038907
节气	-0.130418	-0.069053	-0.066695	-0.148098	0.000942	-0.087376	-0.026397	-0.184699
中草药	0.075387	0.037861	0.06697	0.113785	0.271857	0.024335	-0.144011	-0.159377
东方文化	0.081817	0.040611	-0.083103	0.050101	0.202689	0.050573	-0.265132	-0.339921
热门话题	0.035618	-0.099495	0.100005	-0.082705	0.082995	-0.136864	-0.09543	0.058307
中奖公布	-0.158982	-0.126234	-0.080758	-0.152078	-0.09767	-0.019959	0.101126	-0.002958
品牌价值	-0.092075	-0.019535	-0.030104	0.078554	-0.143439	0.041116	-0.04918	0.026625
护肤知识	0.017689	0.004092	0.109806	-0.083366	0.01726	-0.115995	-0.122342	-0.061067
生活方式	-0.090533	-0.073045	-0.037247	-0.040752	0.014622	-0.029715	-0.063122	0.054753
心灵鸡汤	-0.099596	-0.059162	-0.006051	-0.072748	-0.20583	-0.031439	0.049214	0.132171
时尚穿搭	-0.05343	-0.031749	0.036862	-0.014034	-0.075204	-0.041374	-0.03858	0.042738
跨界合作	-0.034321	-0.026463	-0.010177	-0.04586	-0.083189	-0.028147	-0.057952	0.094649
品牌促销	-0.083212	-0.069626	-0.038592	-0.083062	-0.125849	-0.039548	0.460744	-0.107389
测评	0.177027	0.125446	0.188188	0.064363	0.127981	0.029037	-0.07935	-0.097534
美肌课堂	0.045054	-0.003633	-0.0097	-0.003581	0.114613	0.094032	0.067421	0.053052

说明:

从时间来看: 周一有利于提高阅读量、转发量、点赞、留言；周三有利于提高阅读量和留言；周五有利于提高留言和新关注人数；在周四、周六、周日发表公众号文章并不能带来新的关注人数。因此不建议在这几天发表文章。

从内容来看: 影视能带来更多阅读量和收藏量；活动能同时带来新关注人数和取消关注人数，办活动要慎重。热门话题、护肤知识、测评都能带来更多收藏，测评同时还带来了更多转发和留言。中草药、东方文后和美肌也能带来更多留言，但品牌促销、心灵鸡汤等内容对留言数量的提高起到了副作用；中奖公布会带来很多负面影响，可能跟很多人没有中奖有关。

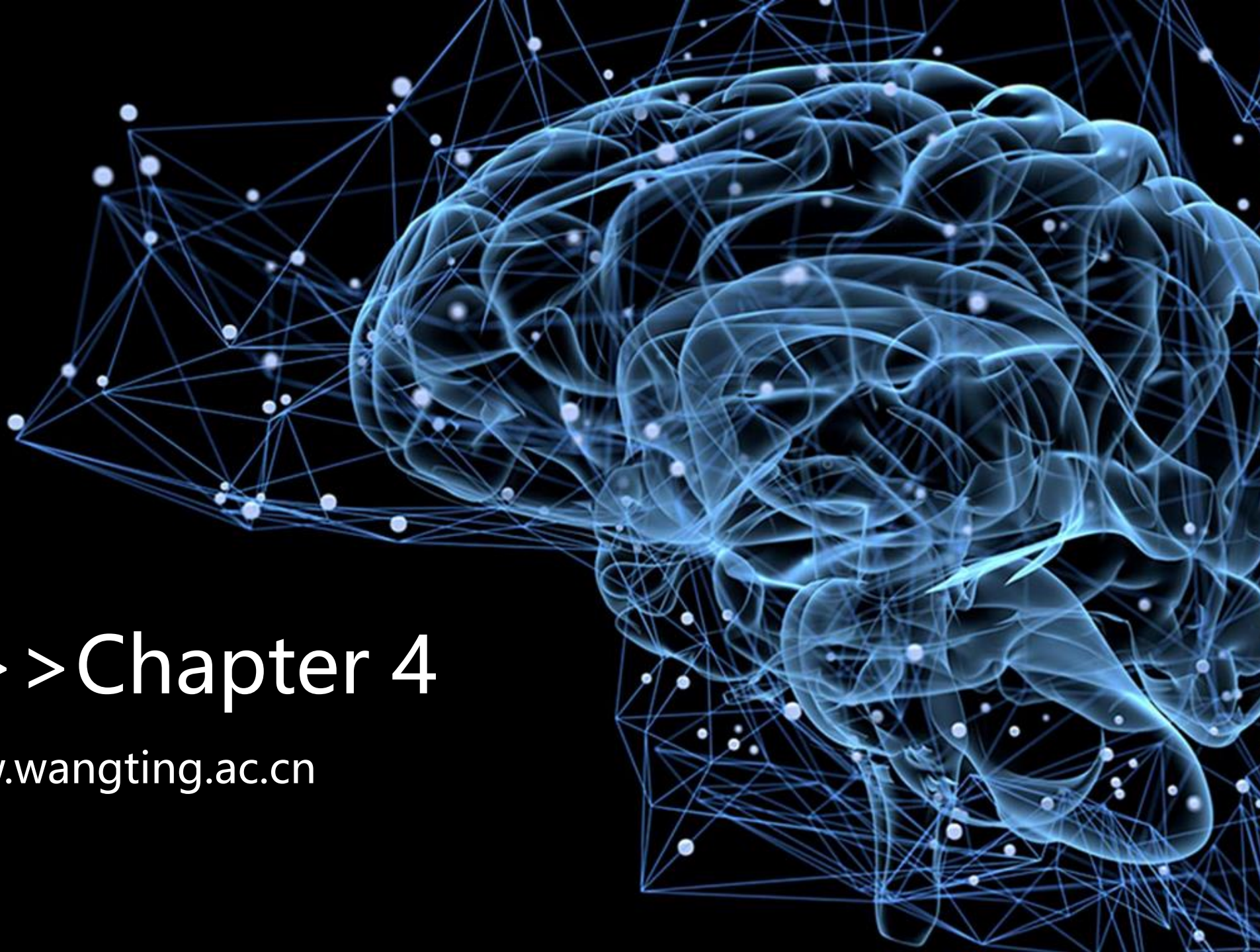
Chapter 3: Data Preprocessing

子产品方面：野果、橙花能增加阅读量，而芯净、四倍蚕丝起到了副作用；橙花同时还能促进转发、点赞、留言和新关注人数的提升，是不错的产品；红景天促进了收藏和点赞；睡莲和芍药促进了留言和新关注人数；莲花促进了留言互动；男士和黑茶产品的推广会使得更多人取消关注，此类文章应该少发。

代言人方面：佟丽娅带来的多方面提升，效果好于刘诗诗；王大陆好于刘烨则说明了我们的客户大多为15-25岁的年轻女性。品牌应着力更多向满足年轻人的需求去发展。

其他方面：点赞、优惠的内容起到了正面作用，而淘宝福利却起到了负面作用。

	阅读量	转发	收藏	点赞	留言	新关注人数	取消关注人数	累计关注人数
野果	0.127025	0.00689	-0.031486	-0.041723	0.091497	-0.12239	0.186397	0.359156
红景天	0.089309	0.098899	0.179622	0.11519	-0.139929	-0.018669	0.080546	0.202386
山茶花	-0.066155	-0.07422	-0.042356	-0.052826	0.031346	-0.044577	0.094424	0.26452
百合	0.012205	-0.040572	-0.022053	0.028329	-0.030882	-0.062635	-0.014268	0.01027
芯净	-0.141874	-0.061834	-0.07093	-0.167497	0.00896	-0.172156	0.086494	-0.147107
男士	0.071727	-0.05896	-0.03272	-0.062954	-0.002008	-0.109107	0.332058	0.08021
四倍蚕丝	-0.150575	-0.066667	-0.076368	-0.207086	-0.093758	-0.073338	-0.111467	-0.142342
睡莲	-0.009089	0.043306	-0.067173	-0.003652	0.245551	0.220248	-0.127266	-0.091921
金银花	-0.059497	-0.024081	-0.032981	-0.061072	-0.000347	-0.008877	-0.054888	0.048927
四倍多萃	0.024639	-0.005162	-0.022053	-0.027366	0.0223	0.029714	0.026282	-0.033904
橙花	0.1443	0.238202	0.009812	0.207804	0.365159	0.144393	-0.006726	0.048058
核桃	-0.002834	-0.001883	-0.010177	-0.021492	-0.017746	-0.037617	0.065608	0.007915
芍药	-0.011339	-0.037848	-0.013783	-0.005088	0.174246	0.125911	-0.072196	0.083336
莲花	0.09566	-0.011259	0.02755	0.088158	0.250136	-0.026248	-0.049318	0.028562
红石榴	-0.002696	-0.044315	-0.037389	-0.025819	-0.037312	-0.053717	0.008577	-0.208472
黑茶	-0.025542	-0.010023	-0.0097	-0.03494	0.023715	-0.035958	0.157114	-0.114036
紫芝	0.046525	0.022076	-0.022053	0.043181	0.041293	-0.054939	-0.112746	-0.163706
金缕梅	-0.054778	-0.03694	-0.038592	-0.075636	-0.068869	-0.001069	-0.054817	-0.144654
防晒	-0.072066	-0.031677	-0.044098	-0.088506	0.075744	-0.084968	-0.114979	-0.191145
光彩立现	0.030118	0.006591	0.00194	0.027778	0.007674	-0.025125	-0.07935	-0.109027
佟丽娅	0.201739	0.297358	-0.030254	0.23515	-0.150242	-0.035193	0.159414	0.184967
刘诗诗	-0.031959	-0.00567	0.21796	0.008891	-0.010863	-0.041859	0.088064	0.09224
王大陆	0.032929	-0.052375	0.565103	-0.027366	-0.00619	-0.054939	0.003111	0.11453
刘烨	0.233037	0.019371	-0.032981	0.048683	0.109266	-0.041374	-0.030426	0.048632
点赞最多	0.357594	0.108471	0.049702	0.216617	0.177725	0.149298	-0.172077	-0.19723
淘宝优惠	0.317135	0.403574	-0.000884	0.330123	-0.151898	-0.02023	0.172241	0.135417
随机抽取	-0.116719	-0.002474	-0.076538	-0.083382	0.208574	-0.023742	-0.172158	-0.125445
淘宝福利	-0.135067	-0.096243	-0.069861	-0.117747	-0.188456	-0.022699	0.257724	-0.06177
问答	-0.023388	-0.01934	0.041511	0.018713	0.402601	-0.02628	-0.146299	-0.108659
品牌H5	-0.090088	-0.057002	-0.050883	-0.109828	-0.137725	0.227546	-0.076961	-0.178951



Next>>Chapter 4

www.wangting.ac.cn