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

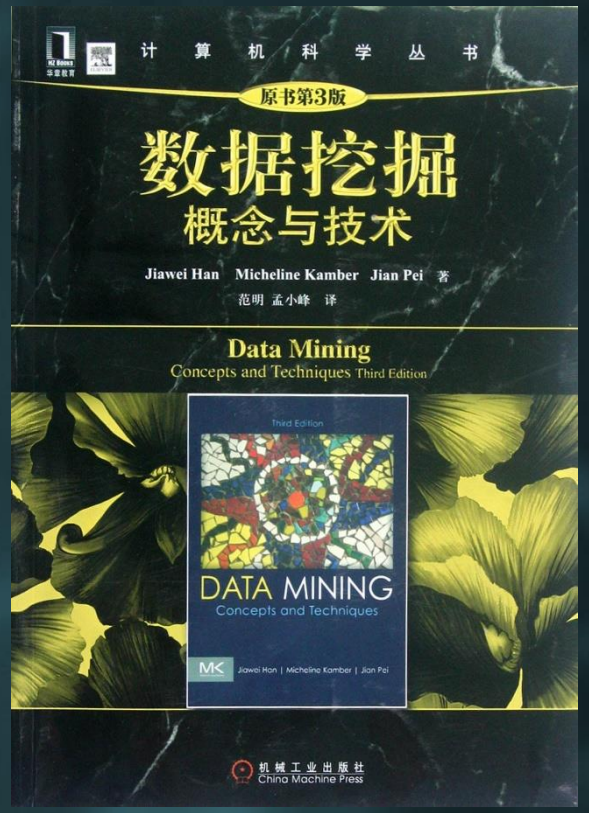
电子信息技术高级工程师

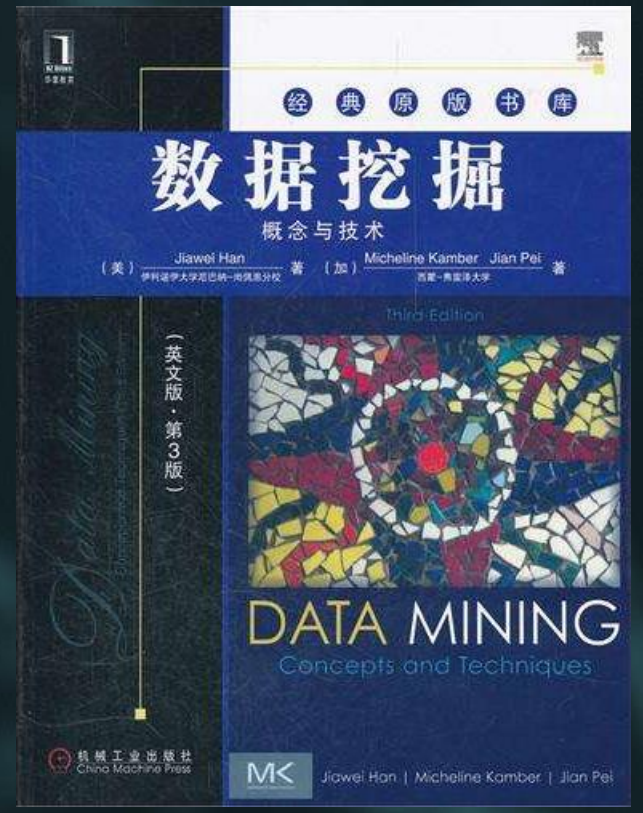
上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任

Chapter 2

Getting to Know Your Data





Chapter 2
Getting to Know Your Data

1. Basic Statistical Descriptions of Data

A. Measuring the Central Tendency（中心性）:

- **Mean（均值）, weighted average（加权平均）**
- **Median（中位数）， median of even numbers（偶数个数的中位数）， median of odd numbers （奇数个数的中位数）**
- **Mode（众数）， unimodal（单峰）， bimodal（双峰）， trimodal（三峰）， …， multimodal（多峰）**
- **Midrange（中列数）：** the average of the largest and smallest values in the set

Chapter 2
Getting to Know Your Data

1. Basic Statistical Descriptions of Data

B. Measuring the Dispersion of Data:

- **Range（极差）**
- **Quartiles（4分位数），percentiles（百分位数），Interquartile Range（四分位差，IQR）**
- **Five-Number Summary（五数概括）：** the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of ***Minimum, Q1, Median, Q3, Maximum***.
- **Outliers（离群点）**
- **Variance（方差）**
- **Standard Deviation（标准差）**

Chapter 2
Getting to Know Your Data

2. Graphic Displays

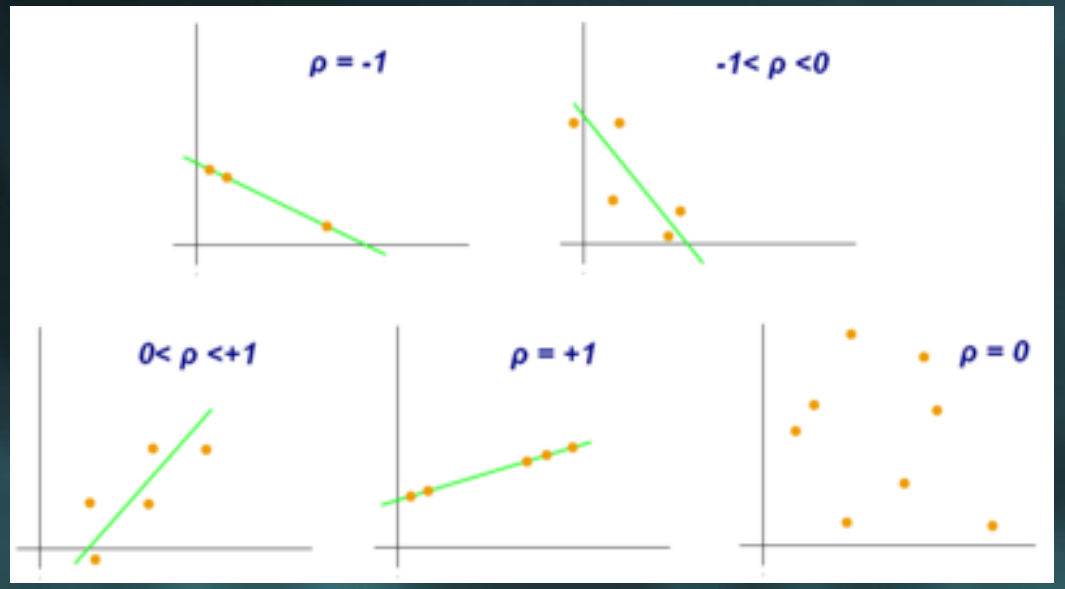
A. Plots

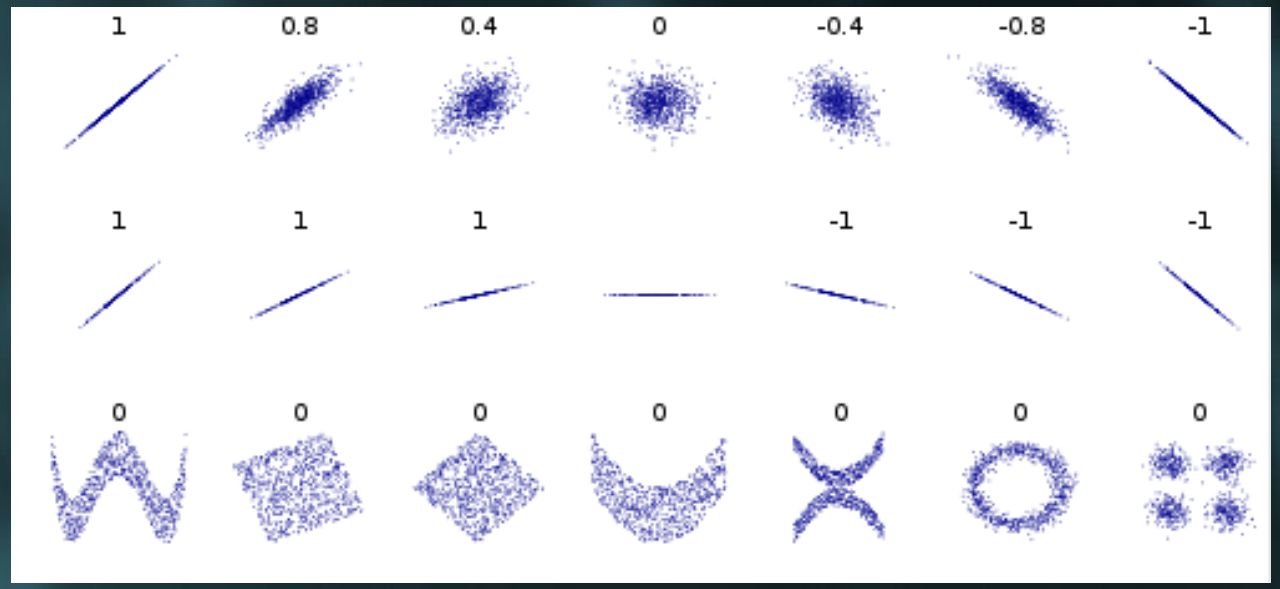
- Boxplots（盒图）
- Histogram （直方图）
- Quantile plot（分位数图）
- Quantile-quantile (q-q) plot （分位数-分位数图）
- Scatter plot（散点图）

Chapter 2
Getting to Know Your Data

2. Graphic Displays

B. Correlation: Pearson Correlation Coefficient





Chapter 2
Getting to Know Your Data

3. Data Similarity

A. Dissimilarity and Distance for Binary Attributes

Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Distance measure for asymmetric binary variables: Jaccard coefficient

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Chapter 2
Getting to Know Your Data

3. Data Similarity

A. Dissimilarity and Distance for Binary Attributes

Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Distance measure for asymmetric binary variables: Jaccard coefficient

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

Chapter 2
Getting to Know Your Data

3. Data Similarity

B. Standardizing Numeric Data

Z-score: $$z = \frac{x - \mu}{\sigma}$$

standardized measure (*Z-score*):

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

where

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

Chapter 2
Getting to Know Your Data

3. Data Similarity

C. Distance on Numeric Data: Minkowski distance

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

- h = 1: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- h = 2: (L_2 norm) Euclidean distance

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) , Chebyshev distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

Chapter 2
Getting to Know Your Data

3. Data Similarity

D. Cosine Similarity

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$

Chapter 2
Getting to Know Your Data

Chapter 2
Getting to Know Your Data

Features

#	Feature	#	Feature	#	Feature	#	Feature	#	Feature
1	United Nations	14	IMF	27	Non Aligned Movement	40	South Asia	53	Latin America
2	OECD	15	Shanghai Cooperation Organization	28	Commonwealth of Independent States	41	Middle Asia		
3	European Union	16	BRICS	29	Treaty on the Non-Proliferation of Nuclear Weapons	42	Middle EAST		
4	Euro Area	17	NATO	30	UNFCCC	43	West Asia		
5	OSCE	18	African Union	31	Asian Infrastructure Investment Bank	44	North Europe		
6	ASEAN	19	League of Arab States	32	African Development Bank	45	East Europe		
7	NAFTA	20	Organization of American States	33	Inter-American Development Bank	46	Middle Europe		
8	G7	21	Commonwealth of Nations	34	Asian Development Bank	47	South Europe		
9	G20	22	Pacific Economic Cooperation Council	35	Comprehensive Progressive Trans-Pacific Partnership	48	West Europe		
10	APEC	23	G77	36	UNESCO	49	Oceania		
11	OPEC	24	South Asian Association For Regional Cooperation	37	WHO	50	North Africa		
12	OAPEC	25	Community of Latin American and Caribbean States	38	East Asia	51	Sub-Saharan Africa		
13	WTO	26	Olympic	39	South East Asia	52	North America		

Chapter 2
Getting to Know Your Data

Jaccard Similarity

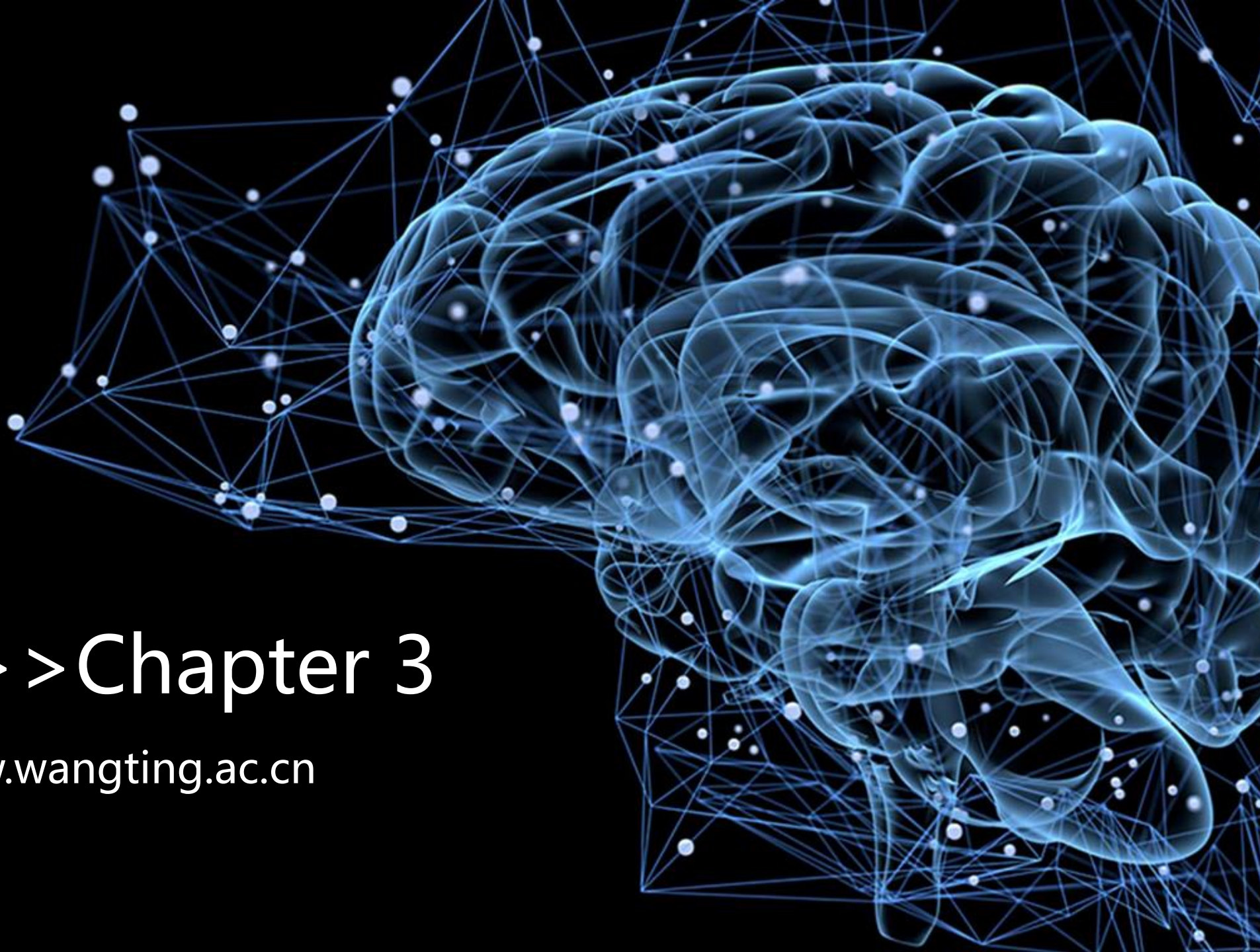
Similarity Analysis

Similarities in Political Entities

Country/Region 1	Country/Region 2	Jaccard Similarity
USA	France	0.64
USA	India	0.344828
USA	Canada	0.875
USA	Spain	0.6
USA	Colombia	0.44
USA	United Kingdom	0.64
USA	Mexico	0.666667
USA	Japan	0.695652
USA	Lithuania	0.458333
USA	Belgium	0.6
USA	Luxembourg	0.52
USA	Netherlands	0.6
USA	Slovenia	0.541667
USA	Nigeria	0.346154
USA	Rwanda	0.346154
USA	Senegal	0.346154
USA	Jordan	0.285714
USA	Argentina	0.48
USA	Ukraine	0.409091

Data Description:

Totally: 195 countries



Next>>Chapter 3

www.wangting.ac.cn