Technology

Science

Information

Networks

Computing

Lecturer: Ting Wang (王挺)

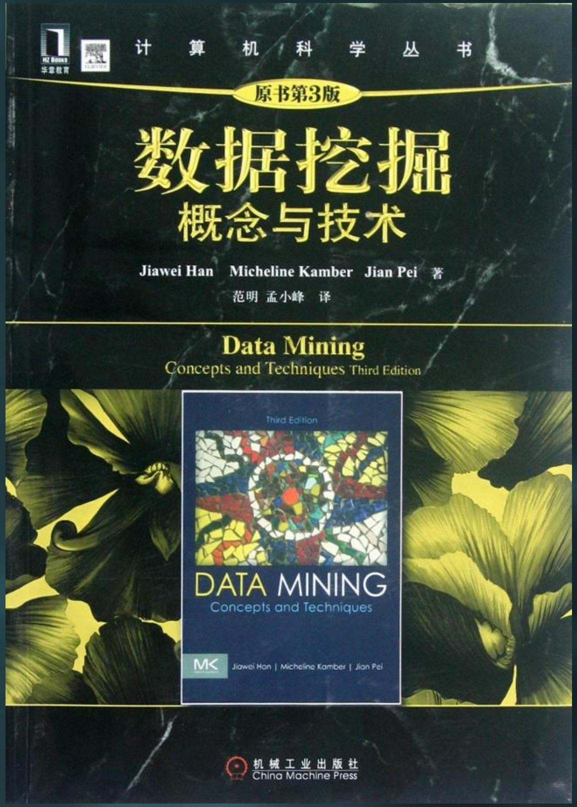
利物浦大学计算机博士

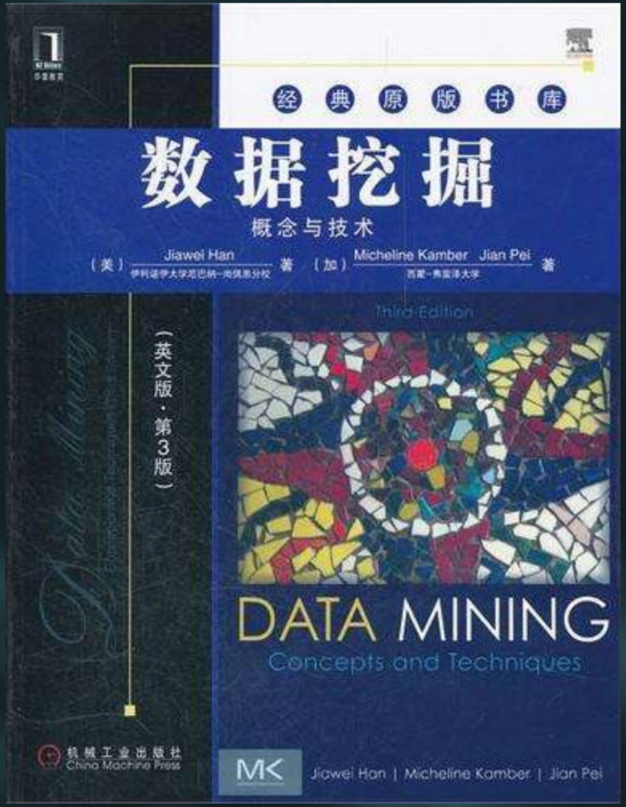
清华大学计算机博士后

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 11

Outliers

Chapter 11 Outliers

1. Outlier

Assume that a given statistical process is used to generate a set of data objects. An **outlier** is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.

Types of outliers:

- global outliers,
- contextual outliers,
- collective outliers.

An object may be more than one type of outlier.

Chapter 11 Outliers

2. Outlier detection

(whether the expert-provided labels are given to the data)

- supervised method
- semi-supervised method
- unsupervised method

(assumptions regarding normal objects versus outliers)

- statistical methods
- proximity-based methods
- clustering-based methods

Chapter 11 Outliers

3. Supervised outlier detection

Modeling outlier detection as a classification problem

- Samples examined by domain experts used for training & testing

Chapter 11 Outliers

4. Unsupervised outlier detection

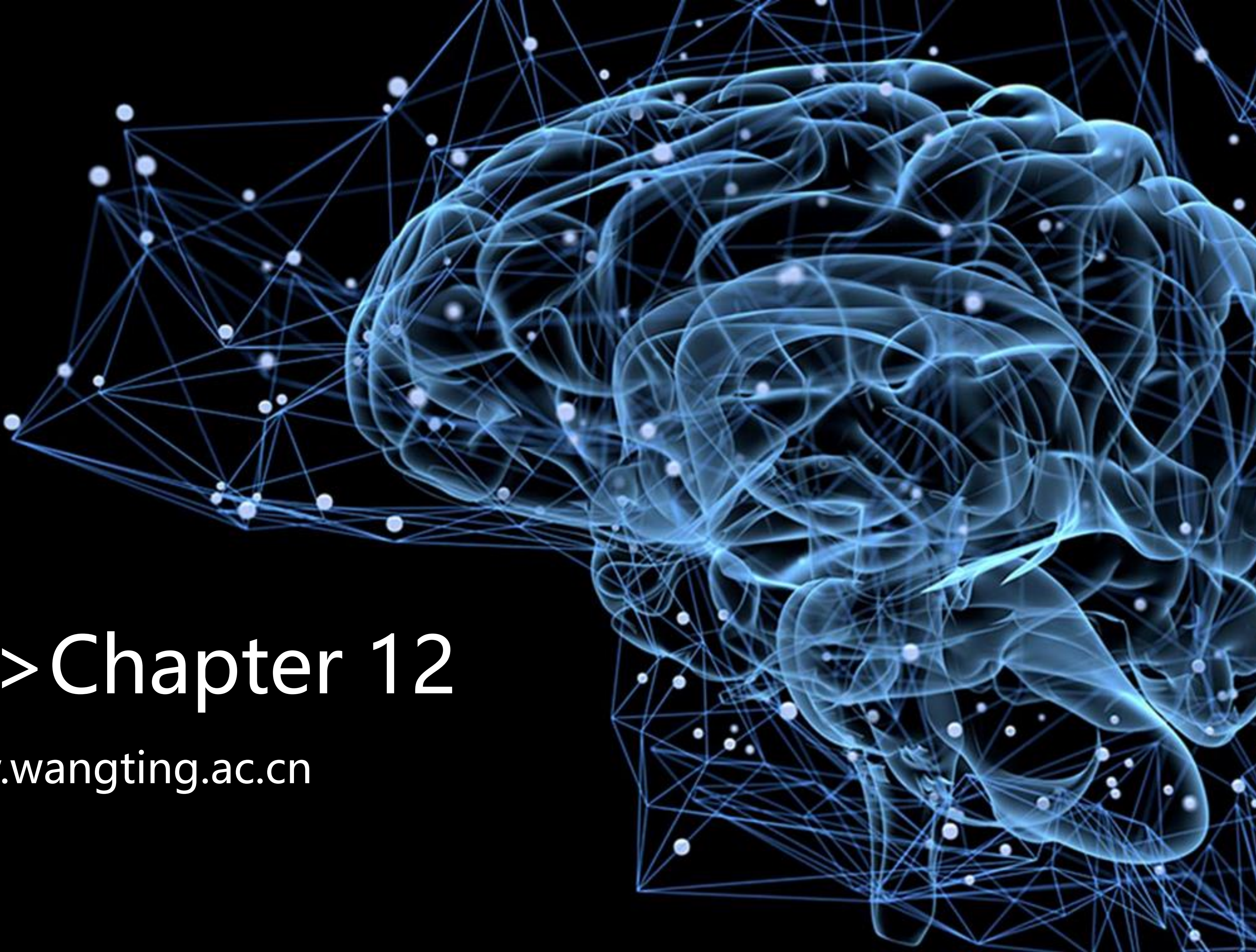
Find clusters, then outliers: not belonging to any cluster

- Problem 1: Hard to distinguish noise from outliers
- Problem 2: Costly since first clustering: but far less outliers than normal objects

Chapter 11 Outliers

5. Semi-supervised outlier detection

- If some labeled normal objects are available
    - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
    - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available, a small number of labeled outliers many not cover the possible outliers well
    - To improve the quality of outlier detection, one can get help from models for normal objects learned from unsupervised methods



Next>>Chapter 12

www.wangting.ac.cn