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

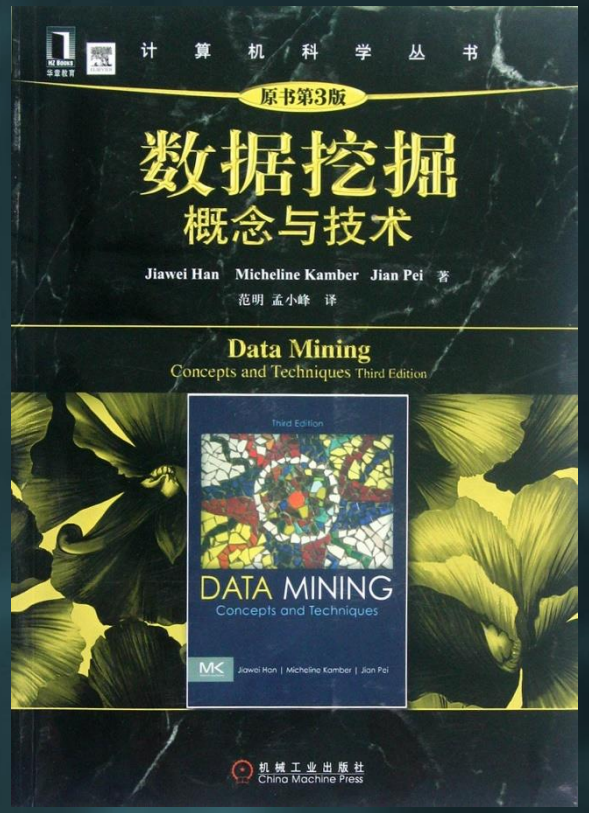
电子信息技术高级工程师

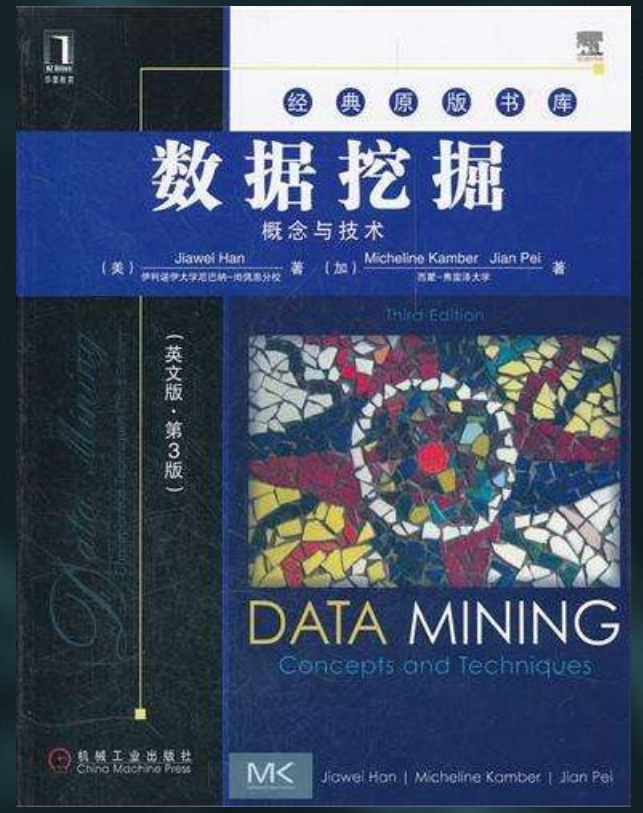
上海外国语大学网络与新媒体副教授

浙江清华长三角研究院海纳认知与智能研究中心主任

Chapter 4

Data Warehousing and OLAP





Chapter 4:
Data Warehousing and OLAP

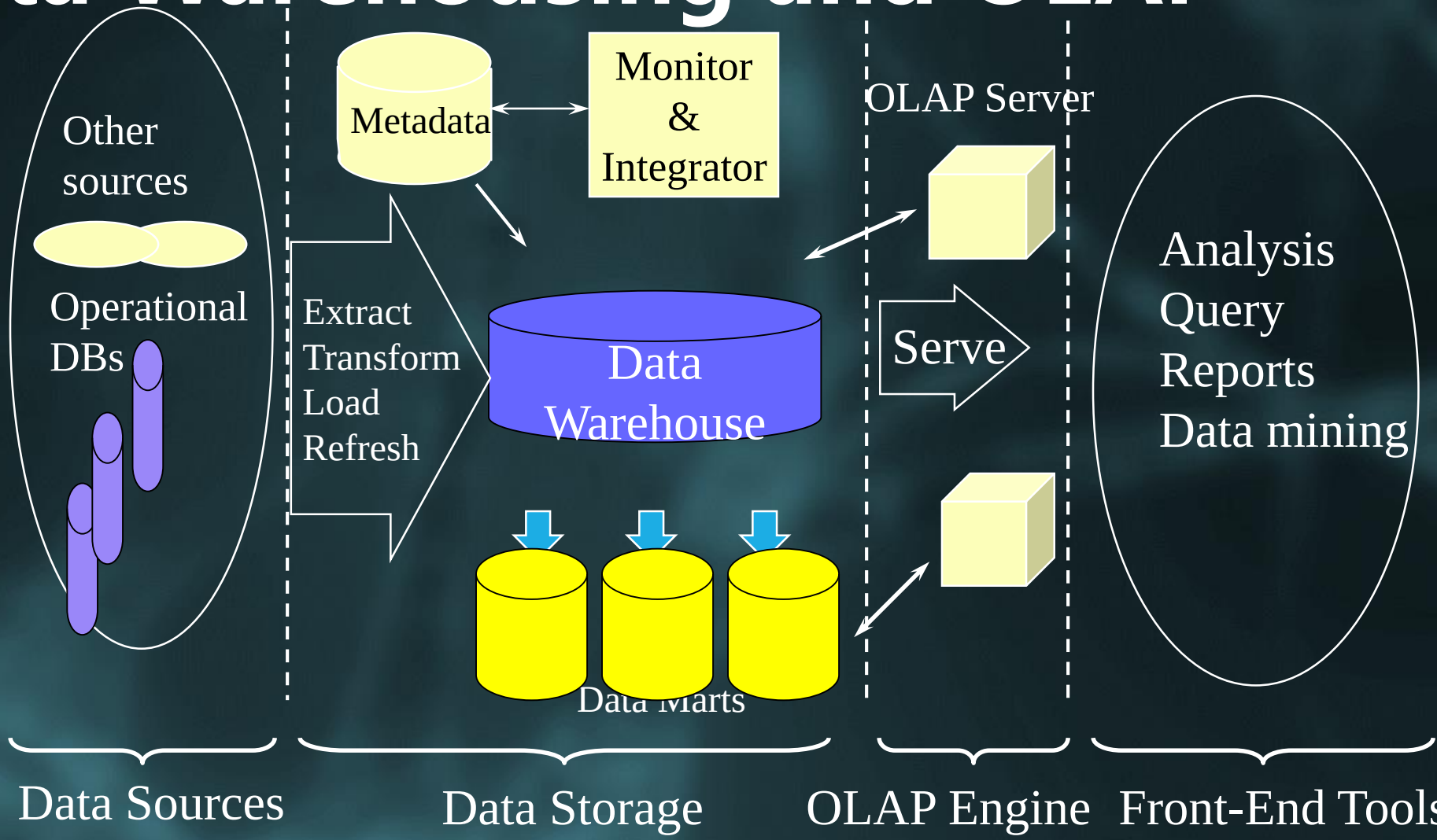
1. Data warehousing: A multi-dimensional model of a data warehouse

- A data cube consists of *dimensions* & *measures*
- Star schema, snowflake schema, fact constellations
- OLAP operations: drilling, rolling, slicing, dicing and pivoting

2. Data Warehouse Architecture, Design, and Usage

- Multi-tiered architecture
- Business analysis design framework
- Information processing, analytical processing, data mining, OLAM (Online Analytical Mining)

Chapter 4: Data Warehousing and OLAP



Chapter 4: Data Warehousing and OLAP

3. The difference between OLTP and OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Chapter 4:
Data Warehousing and OLAP

4. Data Cube

A data cube allows data to be modeled and viewed in multiple dimensions

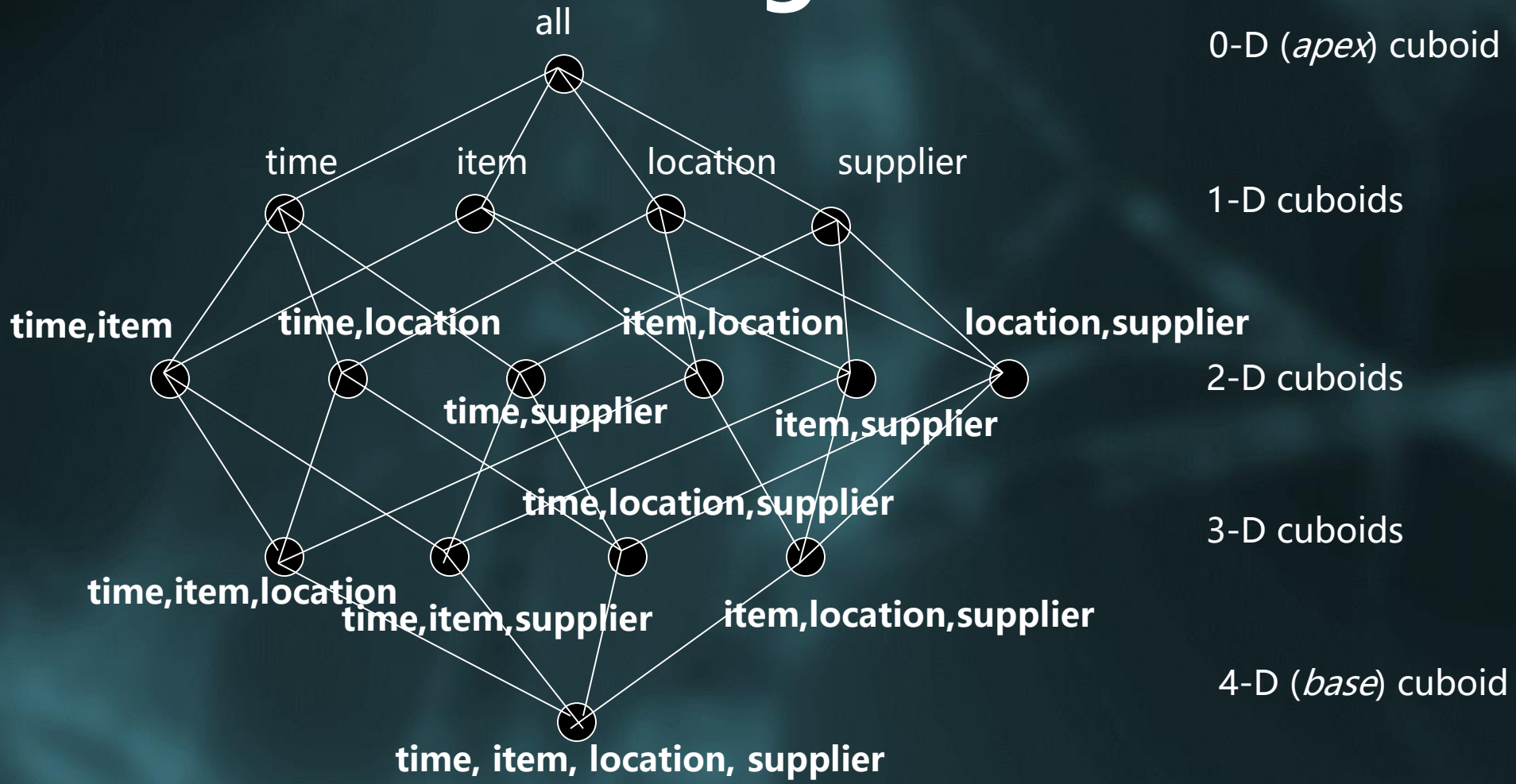
Schema

- Star schema(星型)
- Snowflake schema(雪花)
- Galaxy schema(星系)

Measure

- Distributive(分布性)
- Algebraic(代数)
- Holistic(整体)

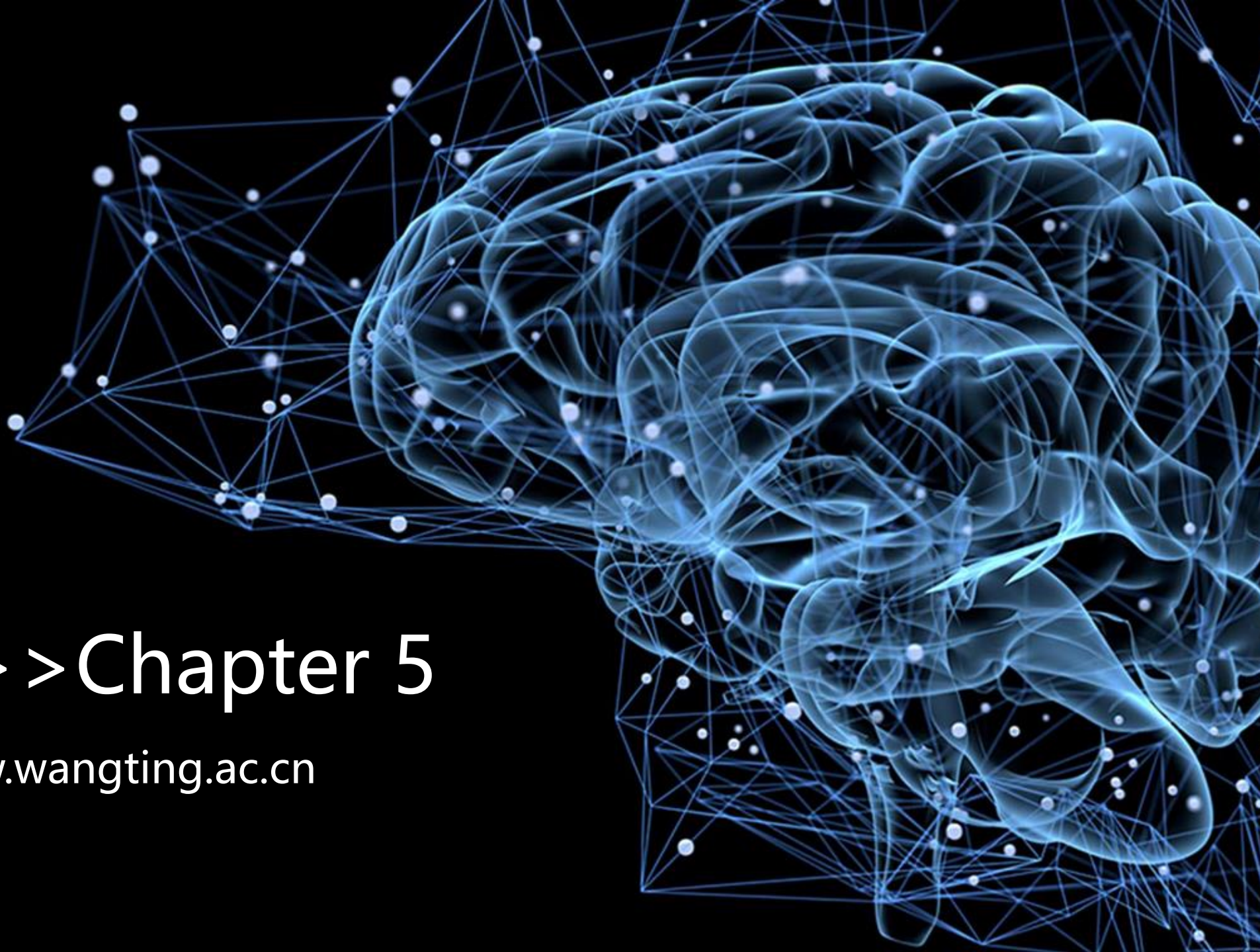
Chapter 4: Data Warehousing and OLAP



Chapter 4: Data Warehousing and OLAP

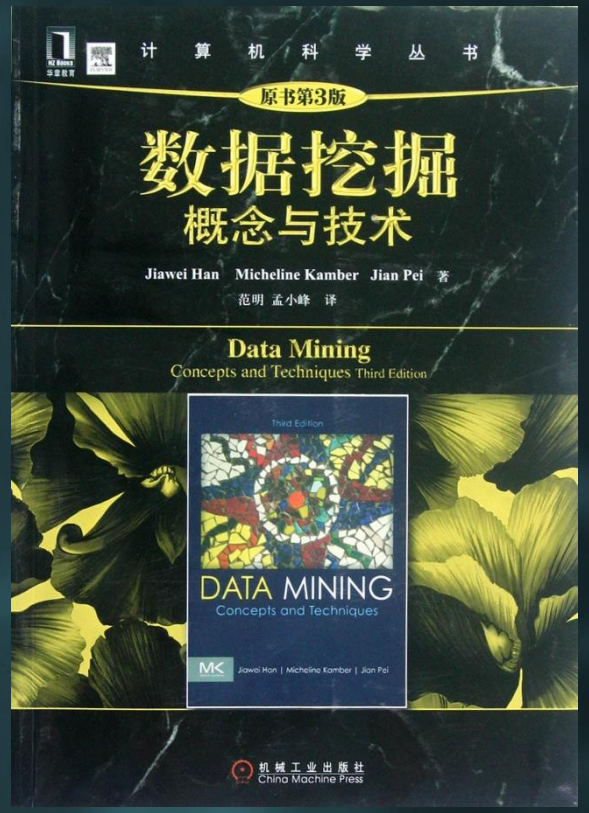
5. Attribute-Oriented Induction（面向属性的归纳)

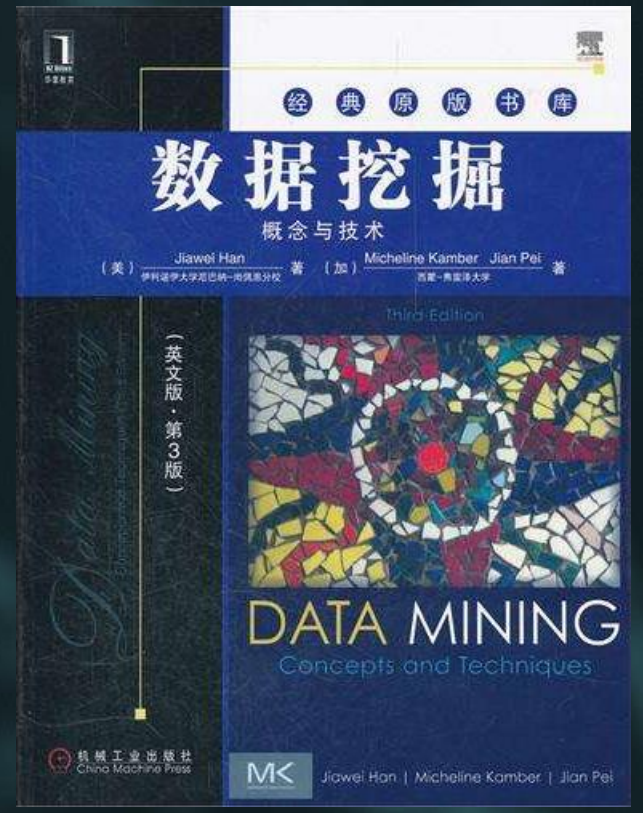
- **How it is done?**
  - Collect the task-relevant data (*initial relation*) using a relational database query
  - Perform generalization by attribute removal or attribute generalization
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts
  - Interaction with users for knowledge presentation
- **Basic Principles**
  - Data focusing
  - Attribute-removal
  - Attribute-generalization
  - Attribute-threshold control
  - Generalized relation threshold control
- **Basic Algorithm**
- **Attribute-Oriented Induction (AOI) vs. Cube-Based OLAP**



Next>>Chapter 5

www.wangting.ac.cn





Chapter 5

Chapter 5: Data Cube Technology

Chapter 5: Data Cube Technology

1. Iceberg Cube

- Computing only the cuboid cells whose measure satisfies the iceberg condition 仅计算满足冰山选择条件的cuboid cells
- Only a small portion of cells may be "above the water" in a sparse cube 稀疏立方体中只有一小部分 "高于水面"
- Avoid explosive growth: A cube with 100 dimensions 避免爆炸性增长:100维的立方体

Chapter 5: Data Cube Technology

2. Methods of Data Cube Computation

- Multi-Way Array Aggregation(多路数组聚合)
- BUC(Bottom-Up Computation, top-down)
- Star-Cubing(Computing Iceberg Cubes by Top-Down and Bottom-Up Integration)
- High-Dimensional OLAP(Semi-Online Computational Model)

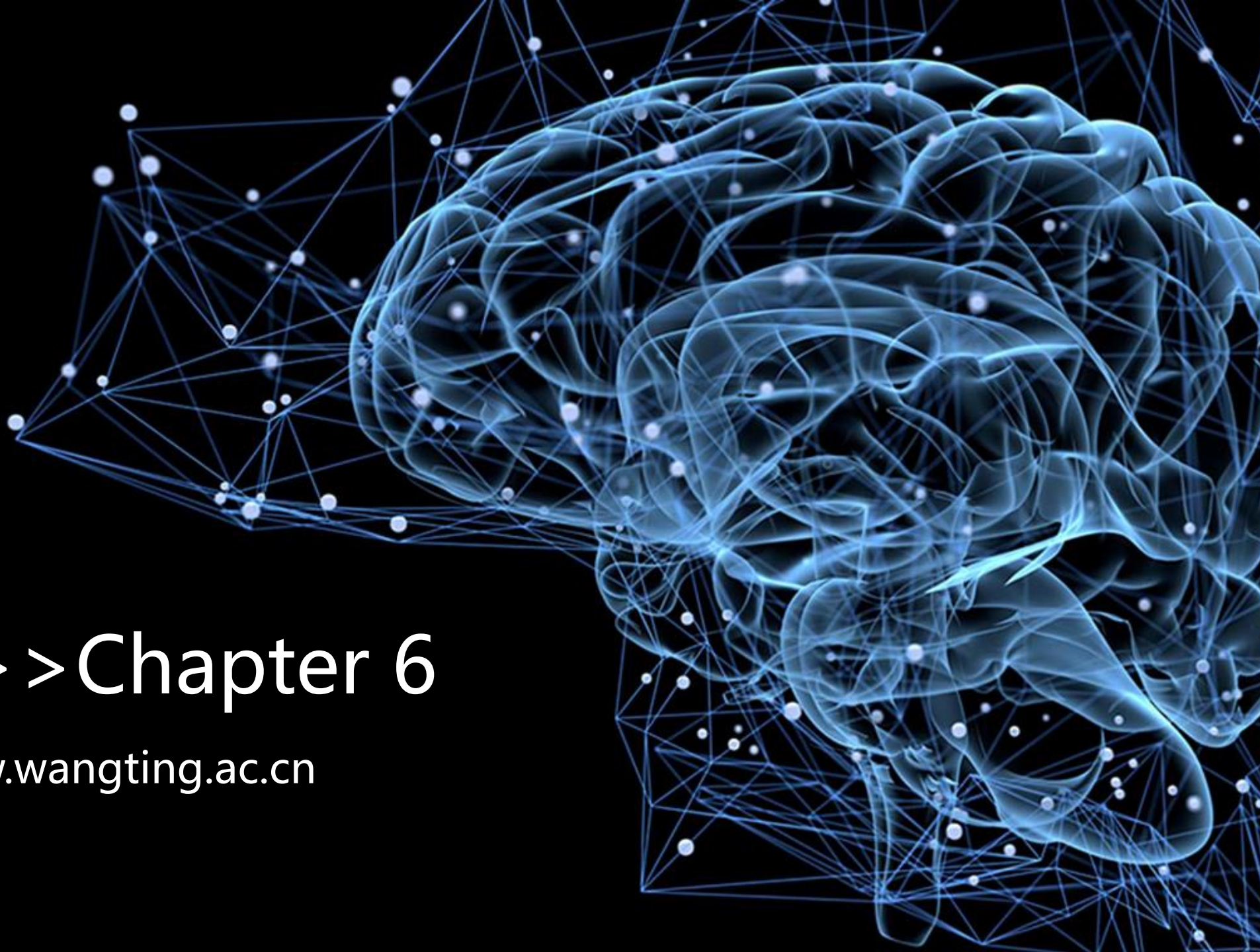
Chapter 5: Data Cube Technology

3. Sampling cube

A data cube structure that stores the sample data and their multidimensional aggregates. It supports OLAP on sample data.

4. Ranking cube

It returns only the best k results according to a user-specified preference. The results are returned in ranked order so that the best is at the top.



Next>>Chapter 6

www.wangting.ac.cn