



CS173 Data Mining

数据挖掘

张海鹏 Haipeng Zhang

School of Information Science and Technology

ShanghaiTech University

个人简介



研究方向：
数据挖掘、金融科技

<https://faculty.sist.shanghaitech.edu.cn/zhanghp/>
zhanghp@shanghaitech.edu.cn

上海科技大学
上海市科学技术委员会
上海市科学技术委员会
中国金融期货交易所

助理教授
金融科技评审专家
扬帆计划 科技英才
博士后

■ 教育背景

• 2005-2009



南京大学

软件工程 学士

• 2007.8-12



香港科技大学

计算机科学 交换生








• 2009-2014



美国印第安纳大学

计算机科学 博士

工作经历

- 2010.6-8  国立情报学研究所（东京） 实习研究员（推荐系统）
- 2012.5-8  eBay研究院（硅谷） 实习研究员（电商大数据）
- 2013.2-5  微软研究院（剑桥） 实习研究员（社交媒体数据）
- 2013.6-9  三星美国研究院（硅谷） 实习研究员（智能手机数据）
- 2014-2015  IBM研究院 研究科学家（大数据、金融科技）
- 2015-2018  中国金融期货交易所 博士后（金融科技、监管科技）
- 2018-至今  上海科技大学信息学院 研究员，博导（数据挖掘、金融科技）

Content

- Course arrangement
- Brief intro into data mining
- Recommended (writing) books

Academic Integrity

- The university code on academic integrity
 - <http://openinfo.shanghaitech.edu.cn/xswyhxgzdjy/list.htm>
- We **DON'T** tolerate academic misconducts
 - ✗ 抄袭他人代码或作品, 抄袭网络源代码(Github, CSDN等)或作品
 - ✗ 反编译别人的编译后文件
 - ✗ 散布源代码
 - ✗ 代他人签到
- 提供抄袭和抄袭别人同等处理
- 鼓励讨论, 仅限于算法层面, 不允许讨论与借鉴实现细节

Topics

- Data processing and modeling
- Clustering
- Data visualization
- Text mining
- Geo-temporal data mining
- (Distributed) data analysis and modeling toolkit
- Social network analysis and graph algorithms
- Research and application scenarios in data mining
- Data-driven project, with end-to-end academic guidance
 - read, write, research, present

Schedule (tentative)

教学周	主要内容
1	1.1 数据挖掘与课程介绍; 1.2 另类数据挖掘
2	2.1 课程课题介绍; 2.2 数据, 数据挖掘流程
3	3.1 数据预处理; 3.2 数据后处理, 数据探索分析
4	4.1 测度指标, 推荐系统; 4.2 Introduction-basic analysis tools
5	5.1课题开题; 5.2 聚类: k-means, 层次聚类
6	6.1 DBSCAN, 聚类评估; 清明节放假
7	7.1 有监督学习; 7.2 有监督学习
8	8.1 课题中期; 8.2 课题中期
9	9.1 Introduction-basic analysis tools (prediction and geo clustering); 9.2 分布式大数据分析存储工具
10	10.1 网络(Network)数据分析挖掘; 10.2 Introduction-NetworkX, 网络(Network)数据分析挖掘 (2) HITS
11	11.1 《Science》数据挖掘论文研读; 11.2 考试
12	12.1 课题结题; 12.2 课题结题

Evaluations

- 作业： 15%
 - 涉及知识以及工具的应用以及文献的阅读。
- 随堂： 12%测验， 3% attendance
- 考试： 25%
- 课程项目： 45%
 - 项目于学期初宣布，以小组为单位完成课题。项目贯穿整个课程期间，将撰写报告并汇报项目，分为开题、中期以及结题。
- 课程总分5%以内的bonus
 - 部分 (~4%) 来自课程项目
 - 部分 (~1%) 来自课堂回答

Inquiries and office hours

- Related inquiries
 - Ask on 教学互动平台
 - Email, put **CS173** in titles
 - Office hours
- TA office hours
 - 杜晓聪, duxc2023@shanghaitech.edu.cn 每周一下午15:00-16:00, SIST 1B-103
- Prof. office hours
 - 每周三 14:00-15:00, SIST A404.C (请邮件预约)

DATA MINING

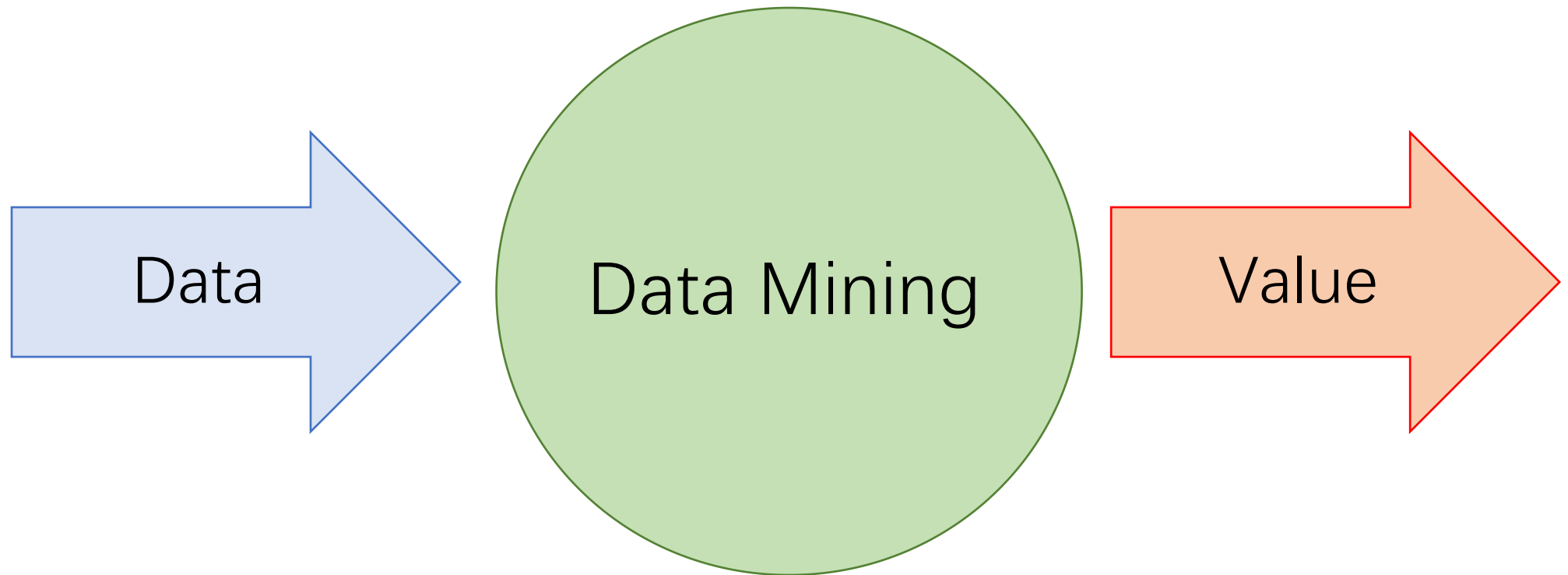
Introduction

What is data mining?
Applications and techniques

- Without **mining**, you are only looking at stones.



Data Mining



What data and techniques are behind
the applications?



- Search hints



- Data: click histories, user demographics (location, language, etc.)
- Techniques: clustering, supervised learning...

Webpage ranking

上海科技大学

上海市浦东新区华夏中路393号201210 (浦东校区) · 上海市徐汇区岳阳路319号8号楼200031 (岳阳路校区) · Copyright © 上海科技大学版权所有沪ICP备13001436号-1.

信息学院

师资队伍 · 研究生培养 · 行政人员 · 学生培养 - ...

机构设置

努力建设一所小规模、高水平、国际化的研究型、创新型大学 ... 科技 ...

学校概况

上海科技大学 (ShanghaiTech University, 简称上科大 ...

校历

Cancel. 上海市浦东新区华夏中路393号201210 (浦东校区) · 上海市 ...

招生

自2017年起, 在理学 (07) 和工学 (08) 学科门类下的物理学 (0702) 、化 ...

大科学中心

科研新闻 · 学术报告 · 研究方向 · 活细胞结构与功能成像等线站工程 ...

[More results from shanghaitech.edu.cn »](#)

<https://zh.wikipedia.org> › zh-hans · [Translate this page](#) ⓘ

上海科技大学- 维基百科，自由的百科全书

上海科技大学 (ShanghaiTech University, 缩写作ShanghaiTech; 筹建时曾用名: 上海高等研究大学), 简称上科大, 是主校区位于中国上海市浦东新区张江高科技园区·中区的 ...
[大事记](#) · [校园设施](#) · [招生](#) · [校园文化](#)

<https://baike.baidu.com> › item › 上海... · [Translate this page](#) ⓘ

上海科技大学_百度百科

上海科技大学 (ShanghaiTech University), 简称上科大 (ShanghaiTech), 位于上海市, 是一



ShanghaiTech University (上海科技大学)

[Website](#)

[Directions](#)

[Save](#)

[Call](#)

University in Shanghai, China

ShanghaiTech University is a research university in Shanghai, China. Its campus is located in the Zhangjiang Hi-Tech Park in Pudong with an academic focus on STEM research. It has five schools and three research institutes and is backed by the Shanghai Municipal Government and Chinese Academy of Sciences.

[Wikipedia](#)

Address: China, Shang Hai Shi, Pudong, 华夏中路393号 邮政编码: 201210

Phone: +86 21 2068 5225

Postgraduates: 1,499

Notable alumni: Yu Deng, Hao Chen, Wenqi Xu, Zhengqian Fu, [MORE](#)

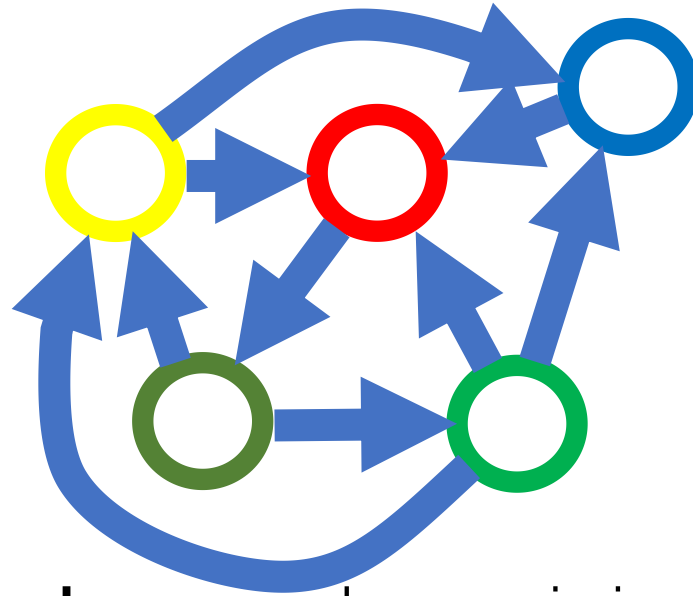
Undergraduates: 1,201

Founder: Chinese Academy of Sciences

Founded: 2013

Campus: 393 Middle Huaxia Road, Pudong, Shanghai, 201210

- Data: webpages and their links



- Techniques: **PageRank** -- a webpage is important if it is pointed to by other important webpages

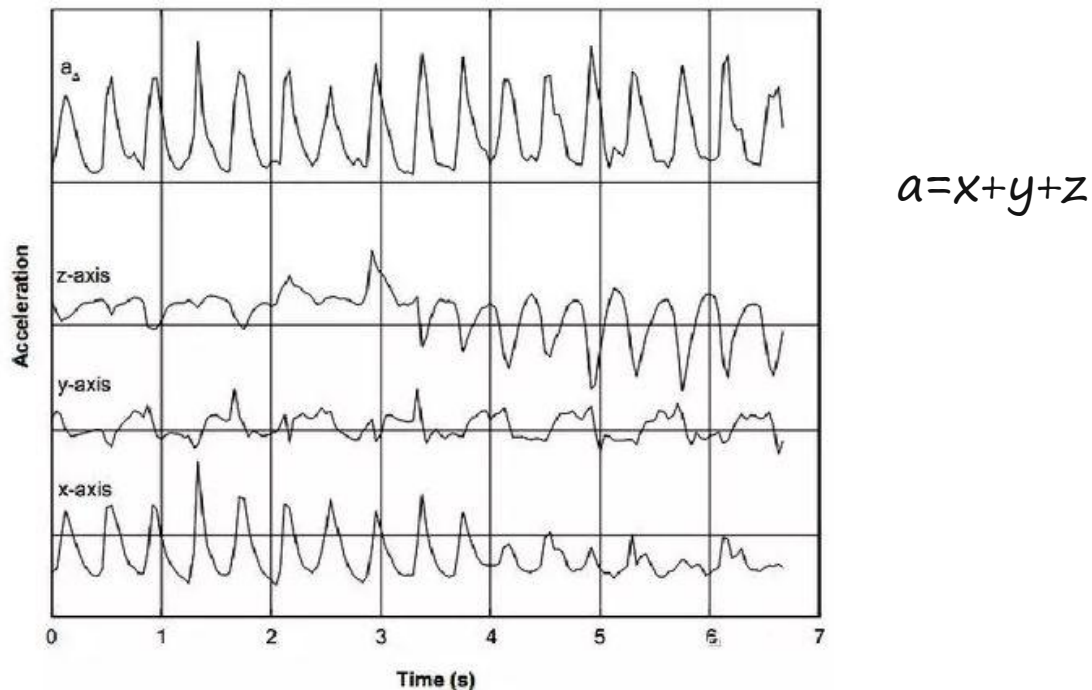
The PageRank Citation Ranking: Bringing Order to the Web (1998)

by Larry Page, Sergey Brin, R. Motwani, T. Winograd

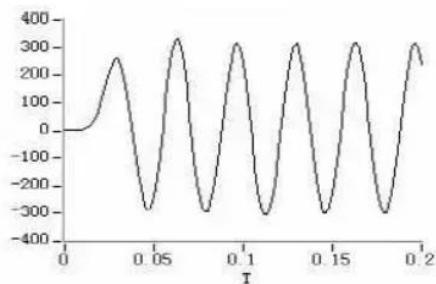
Pedometer 计步器



- 三轴加速度传感器(3-axis accelerometer): 测量在三个不同方向的加速度
- Data: sensor data (x, y, z directions) and labeled data (real steps made)



- Techniques: filtering (signal processing), machine learning



Chinese


▼

↔

English

▼

Enter text



Translation

[Open in Google Translate](#) • [Feedback](#)

Auto driving



The self-driving car's sensors

Just like a person has five senses, Google's self-driving car has a variety of gadgets that detect nearby objects so it can avoid them.

Global Positioning System software
Helps car determine its location.

Position sensor
Located in the wheel hub, this sensor helps determine car's location from wheel rotations.

Radar
Measures speed of cars ahead.

Orientation sensor
Located in car's interior, it acts like the car's inner ear, sensing motion and balance.



Laser
Provides a 360-degree view around the car and helps determine its location.

Microphone
Can detect sirens of approaching emergency vehicles.



Videocameras
With one on each of the car's four corners and another on its roof, they help the car recognize objects around it.



How the car operates

- 1 Any object the vehicle's sensors spot is interpreted by software to determine if it's a pedestrian, cyclist, vehicle or something else.
- 2 Using what it's learned from previous driving, the software makes predictions about what objects will do next.
- 3 The software analyzes the information to decide whether it is safe to accelerate, turn or hit the brakes.

Source: Google
Graphic: Tribune News Service



How the car sees the world

This computerized image is what Google researchers monitoring sensor data see as they ride in the vehicle.

- Other vehicle
- Pedestrian
- Cyclist
- Objects that warrant caution
- A crosswalk, indicating the car needs to stop
- A traffic signal, warning of upcoming railroad tracks
- Path where Google's car intends to go

Recommender system

今日头条

推荐

西瓜视频

热点

直播

图片

科技

娱乐

游戏

体育

懂车帝

财经

搞笑

“从0到1”，习近平反复强调提升这种能力

时政 央视网新闻 · 125评论 · 刚刚

每一句都有力量！中国科学家的奋斗宣言

视频 人民日报 · 303评论 · 4分钟前

执意访台后，捷克参议院主席回国了 结果碰了一鼻子灰

人民日报海外网 · 116评论 · 19分钟前

2020中国民营企业500强榜单

其它 中华工商时报 · 2394评论 · 27分钟前



辽宁阜新：“白菜价”房源折射下的转型困境

房产 中国青年网 · 13评论 · 42分钟前



这个“00后”，是怎么成长为“女枪王”的？

军事 光明网 · 679评论 · 57分钟前

学习强国

搜索站内资讯、视频或用户

搜索

登录后可以保存您的浏览喜好、评论、收藏，
并与APP同步，更可以发布微头条

登录



QQ



微信

24小时热闻



宝能系统频发债 姚振华卷土重来？



“从0到1”，习近平反复强调提升这种能力



家电冤家联手了？格力电器急忙否认

HD 4G 100% 11:42

100% 11:42



关注

推荐

上海



毛衣男 | 男生礼物

搜索

ALL 分类

手机

电动车

垂钓

水族世界

+



¥150

用户_0... 芝麻信用极好

清代人物故事老木雕“文王访贤”，长宽60×35公分

¥100 11人想要

小刚(...)

芝麻信用优秀



包邮 玉化螺 螺盘 小摆件 花纹精美 玉化程度几乎全

¥99 ¥360 6人想要

化石搬运工



包邮 tiny微影 本田 消防

tiny微影 塑料

¥69.50 ¥79 34人想要

全新Tin... 芝麻信用极好



金弹头 马克 感兴趣的



闲鱼



会玩



消息



我的

- Intuition: you like what similar users like -> finding similar users
- Data: behavior data (clicks, purchases) and demographic data (location, age, gender)
- Method: clustering to find similar users and various machine learning methods to make predictions

The data is **complex** and **interconnected**

- Multiple **types** of data: database tables, text, time series, images, videos, graphs, etc
- **Spatial** and **temporal** aspect
- **Interconnected** data of different types:
 - Data generated from a phone: location of the user, friendship information, check-ins to venues, status updates, images through cameras, queries to search engines
 - Very depictive if well aggregated

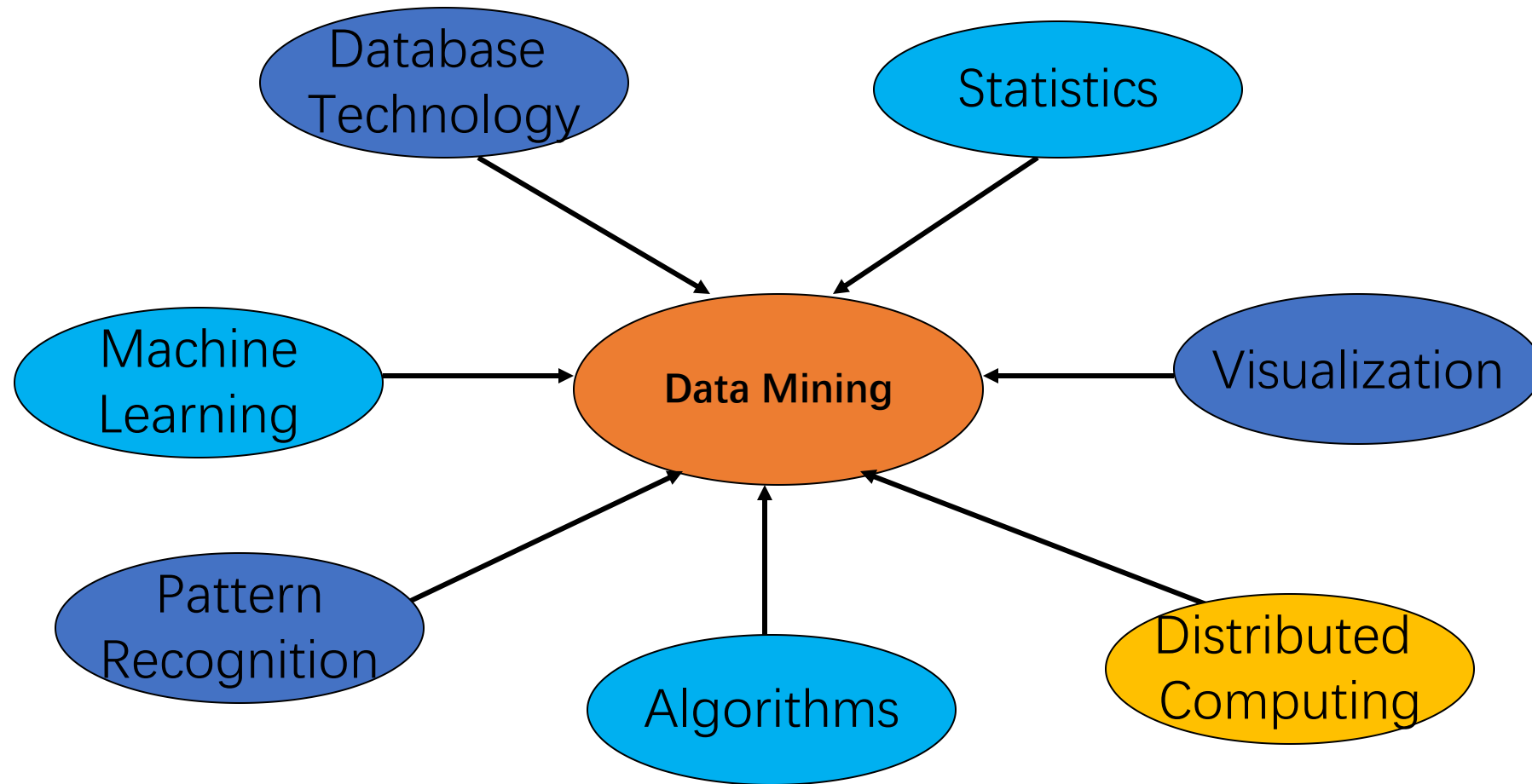
What is data mining again?

- “Data mining is the discovery of **models** for data” (Rajaraman, Ullman)
 - We can have the following types of models
 - Models that **explain** the data (e.g., a single function)
 - Models that **predict** the future data instances.
 - Models that **summarize** the data
 - Models that **extract** the most prominent **features** of the data.
- “Data Mining is the study of **collecting, processing, analyzing, and gaining useful insights** from data” – Charu Aggarwal

Why data mining?

- **Scientific** point of view
 - Scientists are at an unprecedented position where they can collect TB of information
 - Examples: Sensor data, astronomy data, social network data, gene data
 - Analyze such data to get a better understanding of the world and advance science and help people
- **Commercial** point of view
 - As the key competitive advantage for growth
- **Scale** (in data **size** and feature **dimension**)
 - Enormity of data, **curse of dimensionality**
 - The amount and the complexity of data does not allow for traditional processing of the data. We need automated techniques.

Data Mining: Confluence of Multiple Disciplines



New era of data mining

- Boundaries are becoming less clear
 - Today data mining, machine learning, and AI are synonymous. It is assumed that the algorithms should scale.
 - Data is the engine for AI
 - Data Mining touches everything related to data.

Recommended Books

- **Introduction to Data Mining** 数据挖掘导论（原书第2版），Pang-Ning Tan et al.
- **Foundations of Data Science**, Avrim Blum, John Hopcroft, and Ravindran Kannan, free online:
<https://www.cs.cornell.edu/jeh/book.pdf>

Writing in CS:

- [Zobel - Writing for computer science 3rd edition](#) guide on writing for computer science

One last thing:
Data mining can be **artistic** and
answer important questions of
humankind

<https://www.nature.com/articles/d41586-019-03308-7>

<https://www.youtube.com/watch?v=GW4s58u8PZo&t=32s>

<https://www.bilibili.com/video/BV12J411Y7EY/>

