数据挖掘

郑子彬　副教授

中山大学 数据科学与计算机学院

zhzibin@mail.sysu.edu.cn

http://www.inpluslab.com

2016年

课程信息

- 课程名称：《数据挖掘》
  - 浅显易懂地介绍各种数据挖掘技术

- 教材
  - 《数据挖掘导论》, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 人民邮电出版社
  - 《Introduction to Data Mining》， Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN: 9780321321367
  - 相关论文及其他材料

教师信息

- 教师：郑子彬, 副教授（百人计划）
- 助教：吴垚明(ymwu@inpluslab.com)
  叶泳坚(yjye@inpluslab.com)
- Office: 东校区南实验楼D203
- 课程主页：http://dm16.github.io

成绩评定

项目	比例
期末考试	50%
Project	30%
Homework	15%
考勤	5%

期末考试方式：闭卷考试
Project :Kaggle比赛
Homework : 3 times

课程安排

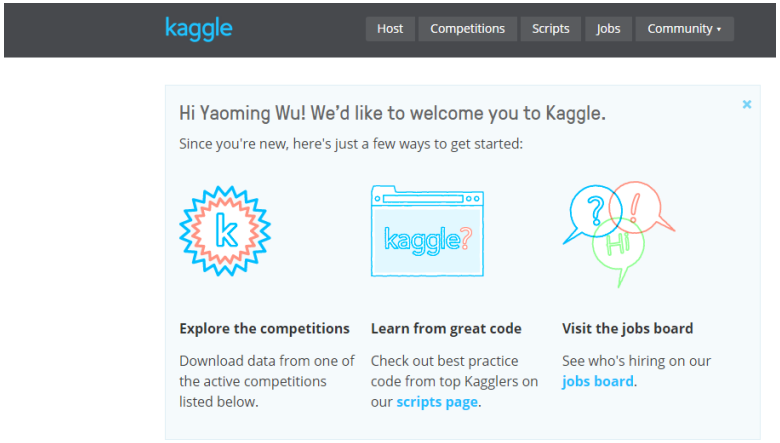
周次	课程
1	Introduction
2	Data
3	Data Exploration
4	Classification
5	
6	
7	
8	Association Analysis
9	

课程安排

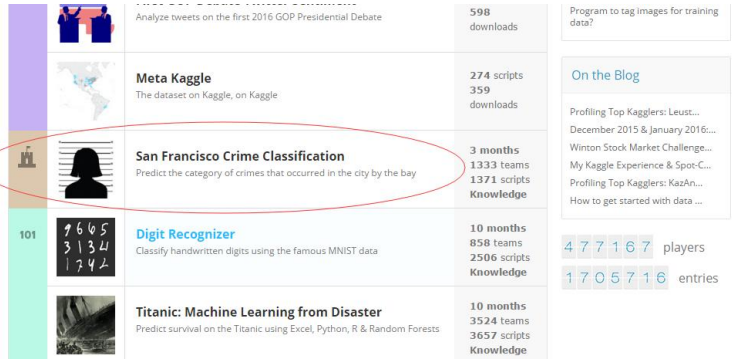
周次	课程
10	Cluster Analysis
11	
12	Collaborative Filtering
13	Time Serial Analysis
14	Graph Mining
15	Computational Advertising
16	Presentation & Review

Project: Crime Classification

- Kaggle: https://www.kaggle.com/



- **San Francisco Crime Classification**



Project: Crime Classification

- This competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.

Project: Crime Classification

- DATA

Knowledge • 1,333 teams

San Francisco Crime Classification

Tue 2 Jun 2015 — Mon 6 Jun 2016 (3 months to go)

Dashboard

- Home
- Data
- Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
- Forum

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
test.csv	.zip (18.75 mb)
sampleSubmission.csv	.zip (2.38 mb)
train.csv	.zip (22.09 mb)

Project: Crime Classification

- Dataset descriptions:
  - **Dates** - timestamp of the crime incident
  - **Category** - category of the crime incident (only in train.csv). **This is the target variable you are going to predict.**
  - **Descript** - detailed description of the crime incident (only in train.csv)
  - **DayOfWeek** - the day of the week
  - **PdDistrict** - name of the Police Department District
  - **Resolution** - how the crime incident was resolved (only in train.csv)
  - **Address** - the approximate street address of the crime incident
  - **X** - Longitude
  - **Y** - Latitude

Project: Crime Classification

- Dataset descriptions:

	A	B	C	D	E	F	G	H	I
1	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
2	2015/5/13 23:53	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258917	37.7745986
3	2015/5/13 23:53	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258917	37.7745986
4	2015/5/13 23:33	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.80041432
5	2015/5/13 23:30	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.4269953	37.80087263
6	2015/5/13 23:30	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.4387376	37.77154117
7	2015/5/13 23:30	LARCENY/THEFT	GRAND THEFT FROM UNLOCKED AUTO	Wednesday	INGLESIDE	NONE	0 Block of TEDDY AV	-122.4032524	37.7134307
8	2015/5/13 23:30	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	INGLESIDE	NONE	AVALON AV / PERU AV	-122.423327	37.72513804
9	2015/5/13 23:30	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	BAYVIEW	NONE	KIRKWOOD AV / DONAHUE ST	-122.3712743	37.72756407
10	2015/5/13 23:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	RICHMOND	NONE	600 Block of 47TH AV	-122.508194	37.77660126
11	2015/5/13 23:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAVENWORTH ST	-122.4190877	37.80780155
12	2015/5/13 22:58	LARCENY/THEFT	PETTY THEFT FROM LOCKED AUTO	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAVENWORTH ST	-122.4190877	37.80780155
13	2015/5/13 22:30	OTHER OFFENSES	MISCELLANEOUS INVESTIGATION	Wednesday	TARAVAL	NONE	0 Block of ESCOLTA WY	-122.4879831	37.73766665
14	2015/5/13 22:30	VANDALISM	MALICIOUS MISCHIEF, VANDALISM OF VEHICLES	Wednesday	TENDERLOIN	NONE	TURK ST / JONES ST	-122.4124143	37.7830038
15	2015/5/13 22:06	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	FILLMORE ST / GEARY BL	-122.4329146	37.78435334
16	2015/5/13 22:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	BAYVIEW	NONE	200 Block of WILLIAMS AV	-122.3977444	37.72993469

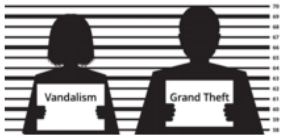
Project: Crime Classification

- Submission Format
  - A csv file with the incident id, all candidate class names, and a probability for each class.

Id	ARSON	ASSAULT	BAD CHECKS	BRIBERY	BURGLARY	DISORDERLY CONDUCT	DRIVING UNDER THE INFLUENCE	DRUG/NARCOTIC	DRUNKENNESS	EMBEZZLEMENT	EXTORTION	FAMILY OFFENSES
0	0.00345693	0.106599	5.31299e-12	0.001958	0.030303	0.000665291	0.00161151	0.0390487	0.00186403	0.000115529	3.78489e-05	0.00134086
1	0.00367083	0.10659	5.31447e-12	0.00195805	0.0303074	0.000678471	0.00161513	0.0401524	0.00186447	0.000115555	3.78631e-05	0.00151509
2	0.00119261	0.0768778	4.51161e-12	0.000806937	0.0363531	0.00160987	0.00178176	0.0317165	0.00377388	0.000302708	6.3377e-05	0.000279141
3	0.00217613	0.106031	5.91491e-12	0.00267633	0.0295841	0.000331426	0.00231467	0.0220826	0.00220504	0.000172331	0.000295773	0.000871006
4	0.00217613	0.106031	5.91491e-12	0.00267633	0.0295841	0.000331426	0.00231467	0.0220826	0.00220504	0.000172331	0.000295773	0.000871006
5	0.0015611	0.0816652	3.70405e-12	0.00130335	0.0336478	0.000857383	0.00272483	0.0216616	0.00450201	0.00020914	0.00015589	0.000952712
6	0.0018385	0.106057	5.91031e-12	0.00267613	0.0295722	0.000313683	0.00230015	0.0203921	0.00220355	0.000172223	0.000295461	0.000618248
7	0.0019181	0.10605	5.91147e-12	0.00267618	0.0295752	0.000318052	0.00230379	0.0208047	0.00220393	0.00017225	0.000295539	0.000673881
8	0.00088918	0.0915652	2.97676e-12	0.00294277	0.0201093	0.00310844	0.0027122	0.0569053	0.00569115	0.000181902	5.65961e-05	0.00102035

Project: Crime Classification

- Leaderboard Score



Knowledge • 1,333 teams

San Francisco Crime Classification

Tue 2 Jun 2015 — Mon 6 Jun 2016 (3 months to go)

Dashboard

Public Leaderboard - San Francisco Crime Classification

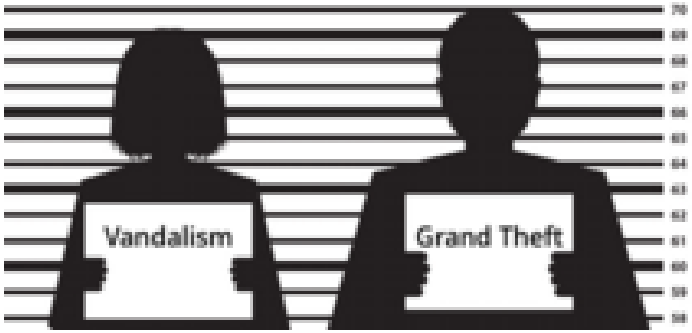
This leaderboard is calculated on all of the test data.

See someone using multiple accounts? Let us know.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	mehran *	2.05079	97	Mon, 11 Jan 2016 15:42:13
2	–	Jghjgfgh	2.06702	12	Sat, 05 Dec 2015 01:44:02
3	–	papadopc	2.11607	111	Tue, 08 Dec 2015 04:54:53 (-32.4d)
4	–	raytrace	2.11638	18	Sat, 05 Sep 2015 19:08:49
5	–	Crime Syndicate	2.13695	20	Sun, 06 Dec 2015 00:04:29
6	–	Wahlen	2.14682	11	Thu, 15 Oct 2015 12:09:40
7	–	elio	2.14714	1	Mon, 17 Aug 2015 21:40:29
8	–	KMen	2.15453	16	Sat, 06 Feb 2016 21:59:40 (-66.4d)
9	–	Markus Loecher	2.18278	11	Mon, 14 Dec 2015 15:22:37

Project: Crime Classification

- Deadline:
  - 6/7/2016 07:59:00 AM UTC+8
- Team Limits: 3 members
- Evaluation:
  - Leaderboard Score (50%)
    - Script
  - Presentation(25%)
    - Algorithm & Model
    - Chanllenges & Solution
    - Outcome
    - 自己实现算法加分，有详细的数据分析加分
    - 注明小组成员贡献百分比
  - Final report (25%)
- More Information:
  - Kaggle: https://www.kaggle.com/c/sf-crime



Project: Crime Classification

- 组队要求：3个小组成员，如有特殊情况（如不够人）可向TA申请
- 组队信息提交到邮箱：dm2016sysu@sina.com
- TA统计好组队信息后将邮件通知各小组组号，Kaggle注册账号请使用DM_组号（如DM_001），评分将根据leaderboard上指定的账号名的分数排名

例子

- 如何找出美国人最喜欢的派？

例子

- 根据销售记录，30寸的派中，苹果派卖的最好

- 对于11寸的派，苹果派只能排4到5名，为什么呢？

例子

- 30寸的派：整个家庭需要都能够接受的口味，苹果派是大家都能接受的，但是不一定是最喜欢的，妥协的结果

- 11寸的派，自己一个人吃，选择自己最喜欢的口味

- **更多的数据 → 获得小数据无法获得的信息**

大数据元年

2012年2月《纽约时报》的一篇专栏中所称，“大数据”时代已经降临，在商业、经济及其他领域中

2012年3月份美国奥巴马政府发布了“大数据研究和发展倡议”

2012年5月，联合国发表名为《大数据促发展：挑战与机遇》的政务白皮书

2012年12月13日被命名为首个“中关村大数据日”

随着一系列标志性事件的发生和建立，人们越发感觉到大数据时代的力量。因此2013年被许多国外媒体和专家称为“大数据元年”。

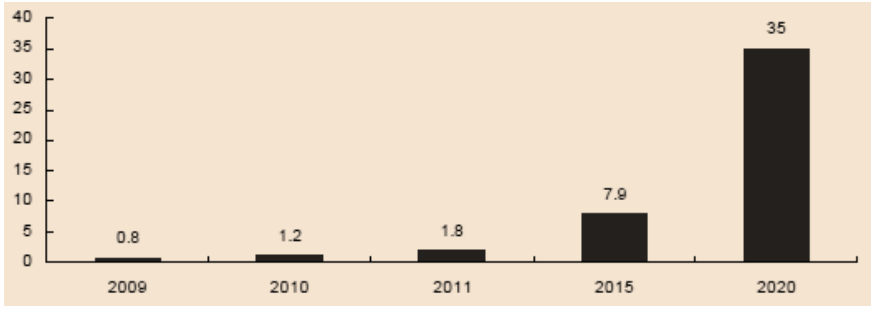
Big Data时代到来

我国网民数量居世界之首，每天产生的数据量也位于世界前列。

淘宝网站	◆单日数据产生量超过5万GB ◆存储量4000万GB
百度公司	◆目前数据总量10亿GB ◆存储网页1万亿页 ◆每天大约要处理60亿次搜索请求
一个8Mbps的摄像头	◆一小时能产生3.6GB的数据 ◆一个城市每月产生的数据达上千万GB
医院	◆一个病人的CT影像数据量达几十GB ◆全国每年需保存的数据达上百亿GB

Big Data时代到来

数据量增加



根据IDC 监测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在2020 年之前会继续保持下去。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量

TB ⇨ PB ⇨ EB ⇨ ZB

从数据库到大数据

“池塘捕鱼”VS“大海捕鱼”　　“鱼”是待处理的数据

- 人类产生的数据早已经远远超越了目前人力所能处理的范畴
- 量级的提升带来的挑战，类比：建筑、管理、系统开发

Big Data时代到来

什么是Big Data

大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合

1. Volume

数据量巨大

全球在2010 年正式进入ZB 时代，IDC预计到2020 年，全球将总共拥有35ZB 的数据量

2. Variety

结构化数据、半结构化数据和非结构化数据

如今的数据类型早已不是单一的文本形式，订单、日志、音频，能力提出了更高的要求

3. Value

沙里淘金，价值密度低

以视频为例，一部一小时的视频，在连续不间断监控过程中，可能有用的数据仅仅只有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”是目前大数据汹涌背景下亟待解决的难题

4. Velocity

实时获取需要的信息

大数据区分于传统数据最显著的特征。如今已是ZB时代，在如此海量的数据面前，处理数据的效率就是企业的生命

大数据时代的机遇和挑战

挑战——大数据技术的运用仍有困难

目前，大数据技术的运用仍存在一些困难与挑战，体现在大数据挖掘的四个环节中。

数据收集

要对来自网络包括物联网和机构信息系统的数据附上时空标志，去伪存真，尽可能收集异源甚至是异构的数据，还可与历史数据对照，多角度验证数据的全面性和可信性。

数据存储

要达到低成本、低能耗、高可靠性目标，要用到冗余配置、分布化和云计算技术，存储时对数据进行分类，通过过滤和去重，减少存储量，并加入便于检索的标签。

数据处理

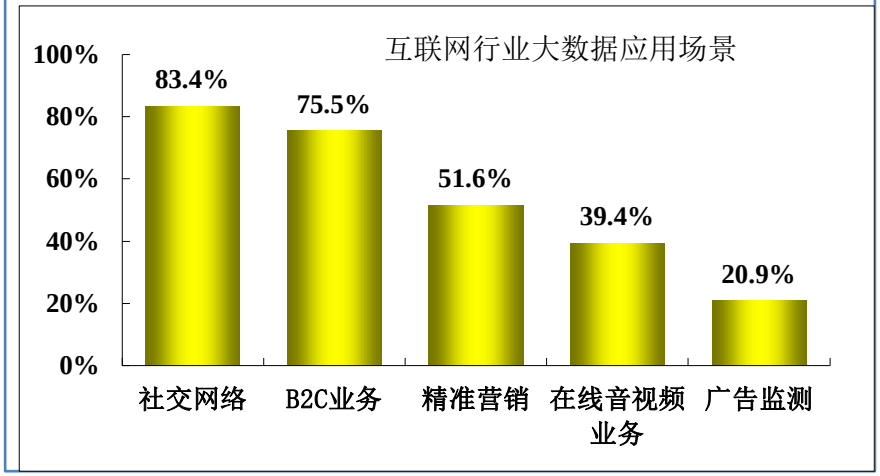
大数据的复杂性使得难以用传统的方法描述与度量，需要将高维图像等多媒体数据降维后度量与处理，利用上下文关联进行语义分析，从大量动态及可能模棱两可的数据中综合信息，并导出可理解的内容。

结果的可视化呈现

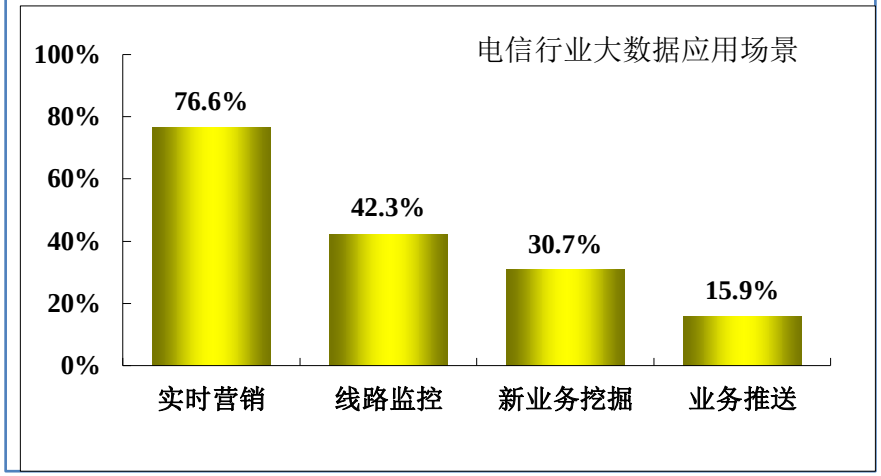
使结果更直观以便于洞察。目前，尽管计算机智能化有了很大进步，但还只能针对小规模、有结构或类结构的数据进行分析，谈不上深层次的数据挖掘，现有的数据挖掘算法在不同行业中难以通用。

大数据潜在应用

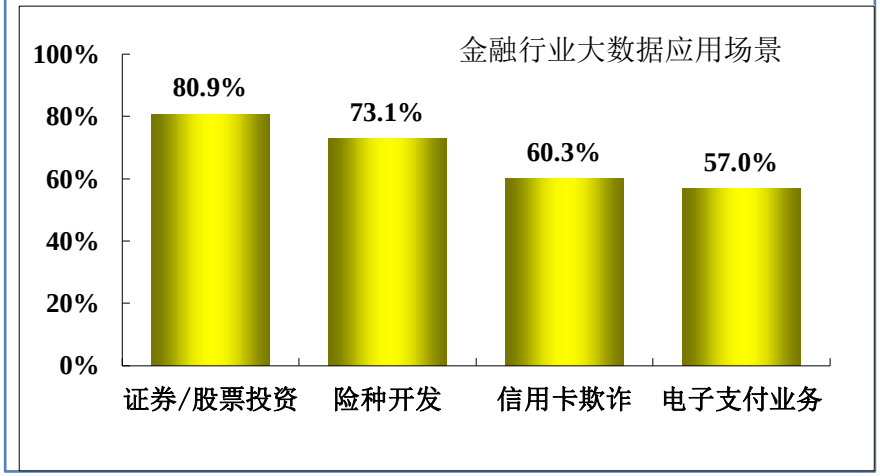
1 互联网行业大数据主要应用在社交和网购方面



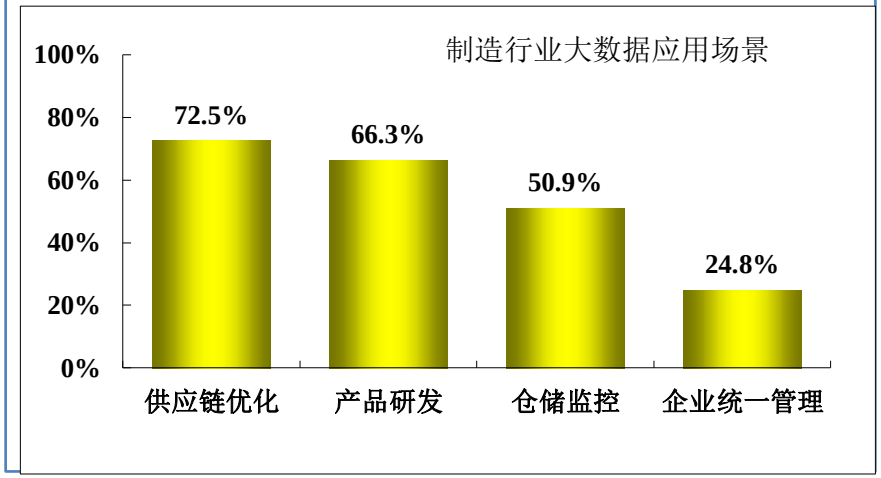
2 结合位置数据、消费数据进行实时营销信息推送是电信行业大数据应用主要场景



3 金融行业大数据应用场景主要集中在投资方面



4 制造行业具有多环节、多地域特色，各个环节的优化是制造行业最关注的大数据应用场景



数据挖掘

- 山西挖矿
  - 前提是有矿，包括煤矿的储藏量，储藏深度，煤的成色
  - 之后是挖矿，要把这些埋在地下的矿挖出来，需要挖矿工，挖矿机，运输机
  - 之后是加工，洗煤，炼丹，等等
  - 最后才是转化为银子
- 数据挖掘
  - 前提是有数据，包括数据储藏量，储藏深度，数据的成色
  - 之后是数据挖掘，要把这些埋藏的数据挖掘出来
  - 之后是把数据可视化输出，指导分析、商业实践
  - 直到这一步，才创造了价值

大数据：一座正在形成的巨型矿山！

大数据 vs 数据挖掘

传统数据挖掘	大数据
• 传统数据挖掘的方法只是从内部数据库数据提取，分析数据	• 大数据则从更多途径，采用更多非结构化的数据
• 处理时间上，传统的对时间要求不高，大多采用集中处理	• 大数据强调的是实时性，数据在线即用，采用分布式处理
• 传统数据挖掘在思维上趋向于使用样本数据集，从中挖掘出残值。	• 大数据着重于总体和全覆盖的理念，收集各种渠道的数据信息，力求全价值。

数据挖掘定义

- **技术角度的定义**

  数据挖掘（Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程

  –近义词：数据融合、数据分析和决策支持等

  –数据源必须是真实的、海量的、含噪声的

  –发现的是用户感兴趣的知识

  –发现的知识要可接受、可理解、可运用

  –并不要求发现放之四海皆准的知识，仅支持特定的发现问题

数据挖掘定义

■ 商业角度的定义

数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性信息

- 一类深层次的数据分析方法
- 可以描述为:按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的有效方法

数据挖掘里程碑

- 在 2015 年二月，DJ Patil成为白宫第一位首位数据科学家
- ……

- 数据挖掘：商业、科学、工程和医药、信用卡交易、股票市场、国家安全、基因组测序、临床试验……

相关领域

- 人工智能
- 机器学习
- 模式识别
- 统计学
- 数据库
- .....

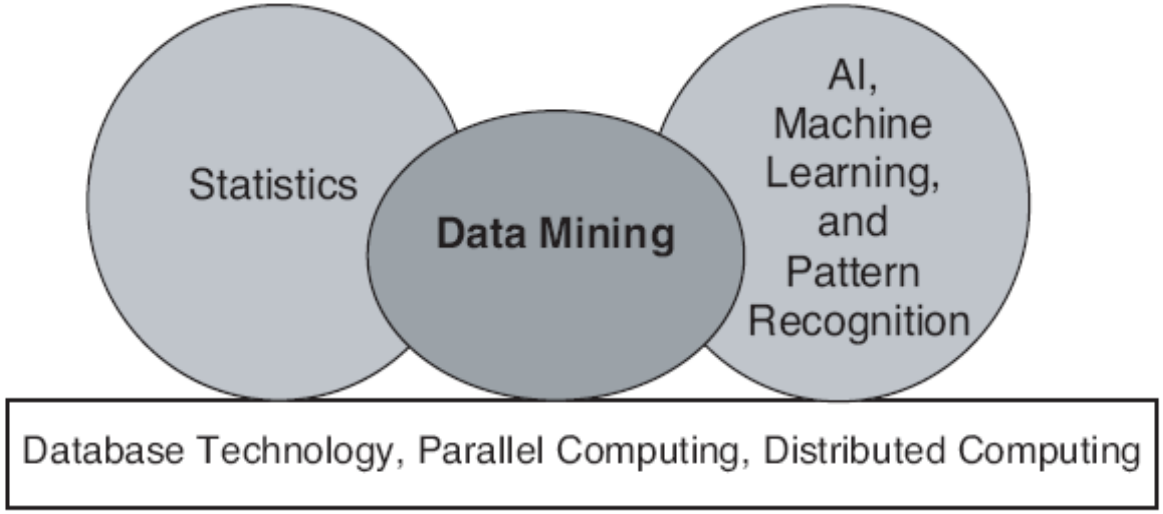
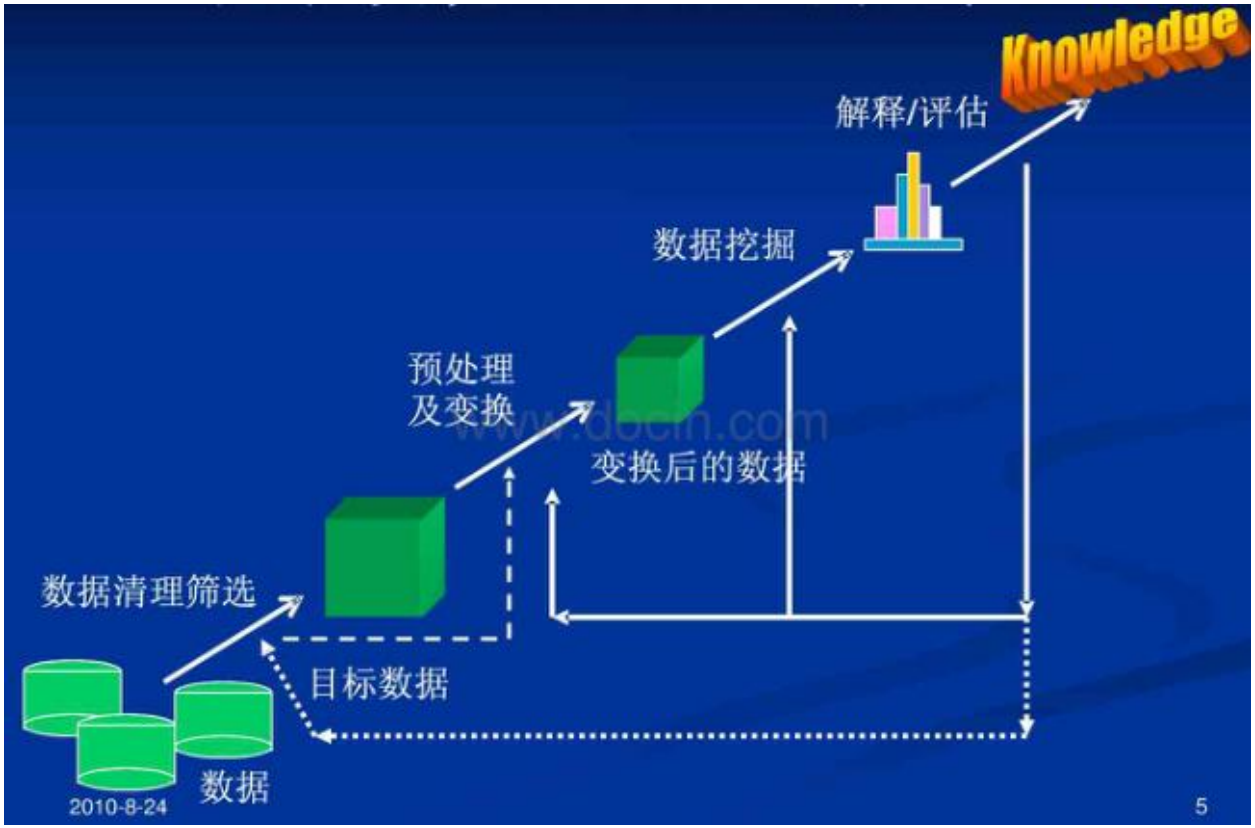


Figure 1.2. Data mining as a confluence of many disciplines.

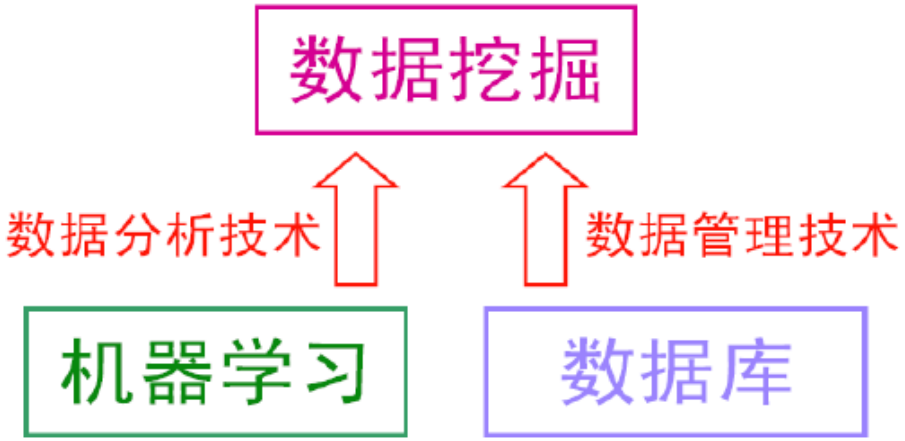
数据挖掘 vs 知识发现（KDD）

- 数据挖掘是KDD中利用算法处理数据的步骤
- 逐渐演变成KDD的同义词



数据挖掘 vs 机器学习

- 机器学习：利用经验来改善计算机系统自身的性能
- 数据挖掘(知识发现)：从海量数据中找出有用的知识
  - 利用机器学习界提供的技术来分析海量数据
  - 利用数据库界提供的技术来管理海量数据



数据挖掘 vs 统计学

- 数据挖掘很多工作由统计方法完成
- 目标相似，许多算法源于数理统计
- 部分统计学家认为数据挖掘是统计学的分支
- 大部分数据挖掘研究人员不这么认为

数据挖掘 vs 传统数据分析方法

- 数据源
  - 数据是海量的
  - 数据有噪声
  - 数据可能非结构化，异构多源
- 传统数据分析方法：假设驱动
  - 给出一个假设，然后通过数据验证
- 数据挖掘：发现驱动
  - 模式从数据中自动提取出来
  - 发现不能靠直觉发现的信息或知识
  - 挖掘出的信息越出乎意料，可能越有价值

相关学术会议

- SIGIR, KDD, ICDM, SDM, CIKM, PAKDD
- WWW, WSDM
- AAAI, IJCAI
- VLDB, SIGMOD, ICDE
- BigData
- ICML, NIPS
- …

相关学术期刊

- IEEE Transactions on Knowledge and Data Engineering(TKDE)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
- ACM Transactions on Intelligent Systems and Technology (TIST)
- ACM Transactions on Information Systems(TOIS)
- IEEE Transactions on Systems, Man, and Cybernetics, Part B
- IEEE Transactions on Neural Network (TNN)
- Knowledge and Information Systems (KAIS)
- Pattern Recognition (PR)

相关比赛

- 阿里天池比赛: http://tianchi.aliyun.com/
- IJCAI: http://ijcai15.org/index.php/repeat-buyers-prediction-competition
- Kaggle: https://www.kaggle.com/
- DataCastle: http://www.pkbigdata.com/
- ImageNet: http://image-net.org/challenges/LSVRC/2015/index
- KDD Cup: https://www.kddcup2015.com/information.html
- Angry Birds AI Competition: http://aibirds.org/

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

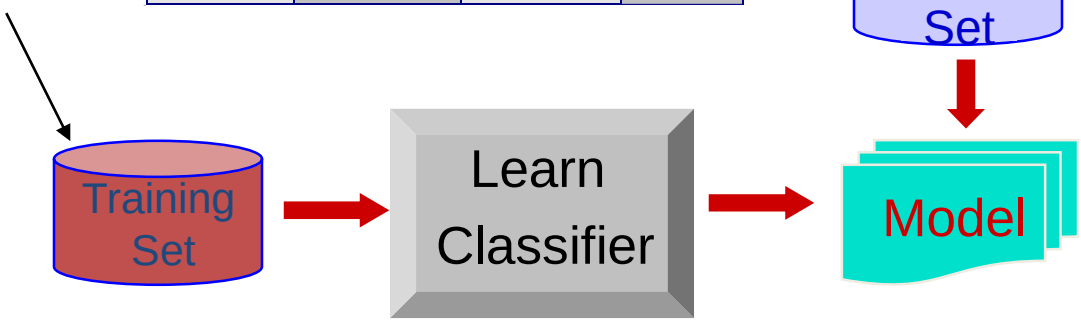
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification: Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Clustering: Definition

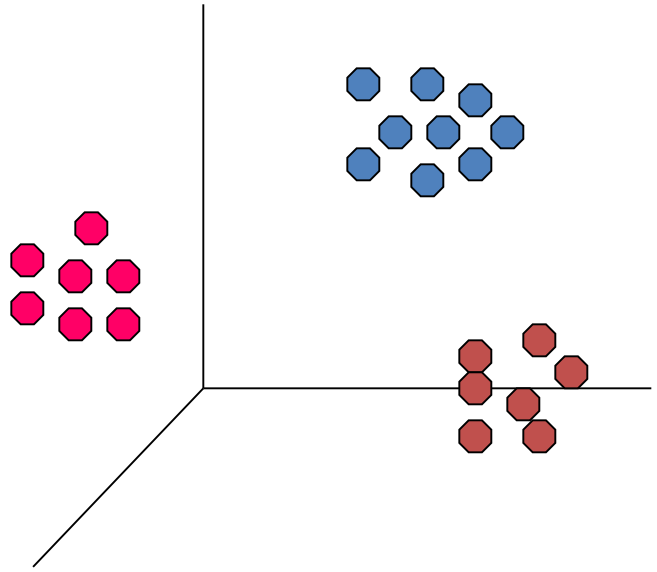
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Examples:

序号	血压	年龄	体重指数	吸烟习惯	序号	血压	年龄	体重指数	吸烟习惯
1	144	39	24.2	0	21	136	36	25.0	0
2	215	47	31.1	1	22	142	50	26.2	1
3	138	45	22.6	0	23	120	39	23.5	0
…	…	…	…	…	…	…	…	…	…
10	154	56	19.3	0	30	175	69	27.4	1

体重指数 = 体重（kg）/身高（m）的平方

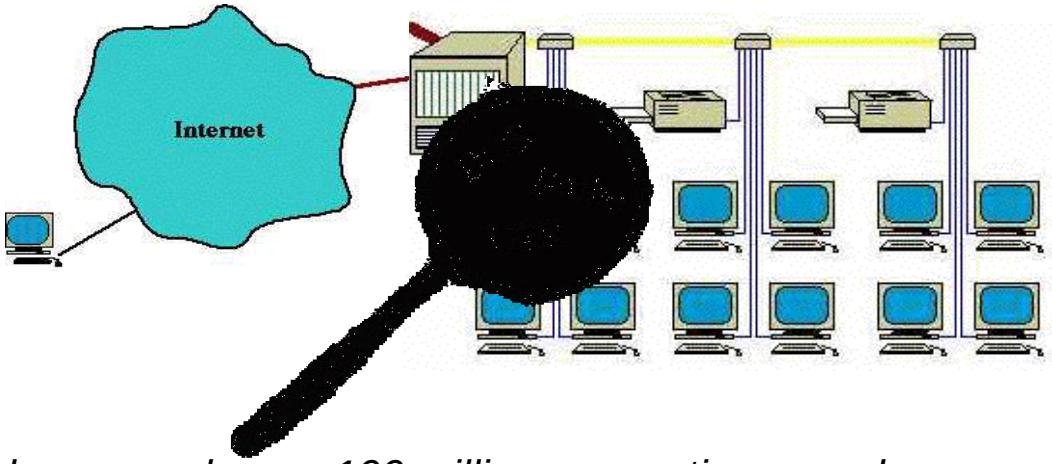
吸烟习惯: 0表示不吸烟，1表示吸烟

建立血压与年龄、体重指数、吸烟习惯之间的回归模型

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

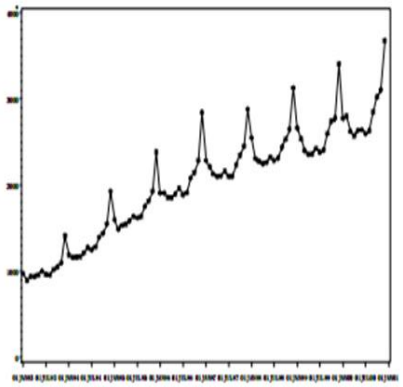
Collaborative Filtering

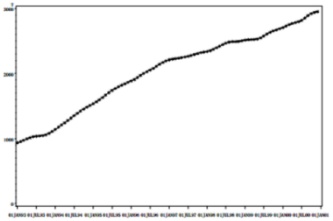
- Recommender System
  - Motivation: Information Overload
  - Solution:
    - Search Engines
    - Recommender Systems
  - Challenges
    - Scalability
    - Cold start
    - Imbalanced dataset
  - Approaches
    - Memory-based approach
    - Model-based approach

		Titles				
		Starship Trooper (A)	Sleepless in Seattle (R)	MI-2 (A)	Matrix (A)	Titanic (R)
Users	Sammy	3	4	3	?	?
	Beatrice	3	4	3	1	1
	Dylan	3	4	3	3	4
	Mathew	4	2	3	2	5
	John	4	3	4	4	4
	Basil	5	1	5	?	?

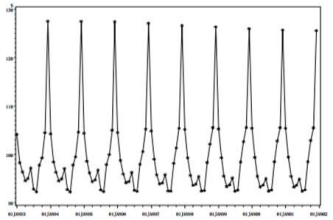
Time Serial Analysis

- 时间序列：用时间排序的一组随机变量
- 时间序列分析：一种根据动态数据揭示系统数据结构和规律的统计方法
- 应用：经济宏观控制、企业经营管理、市场潜量预测、气象、水文、地震、农作物病虫灾害预报、环境污染控制、生态平衡、天文学和海洋学等方面。

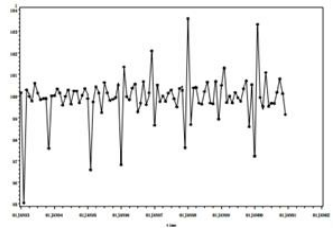




趋势



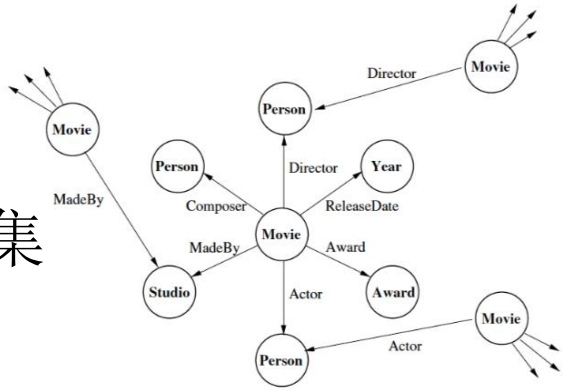
周期（季节）

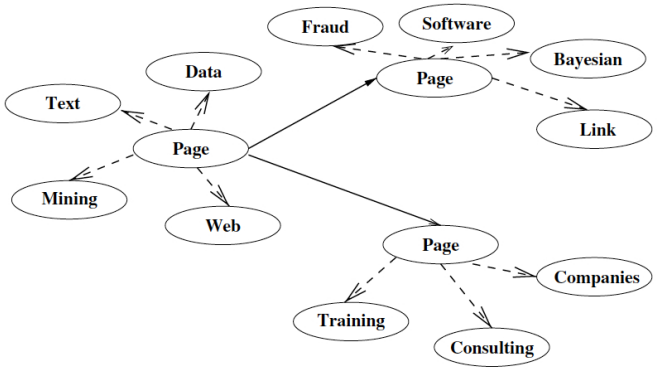


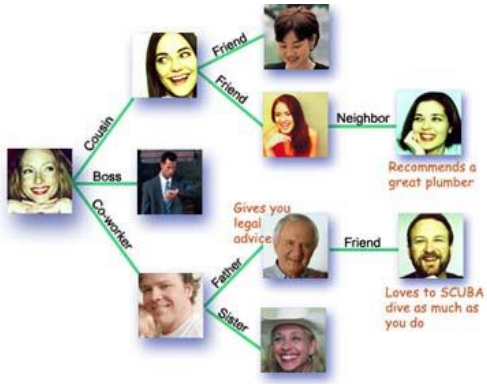
随机

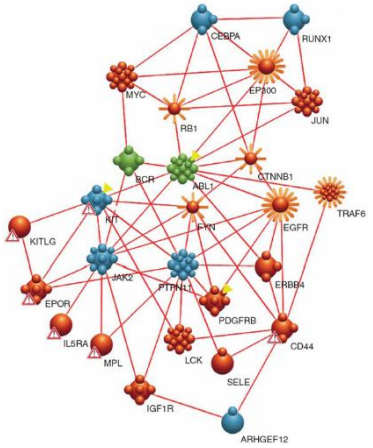
Graph Mining

- 应用领域
  - 网络电影数据集
    - 电影推荐
    - 社区探测
  - 网页数据(Web Data)
    - 网页内容挖掘
    - 网页结构挖掘
    - 网页用途挖掘
  - 社会网络分析
  - 生物信息学









Challenges of Data Mining

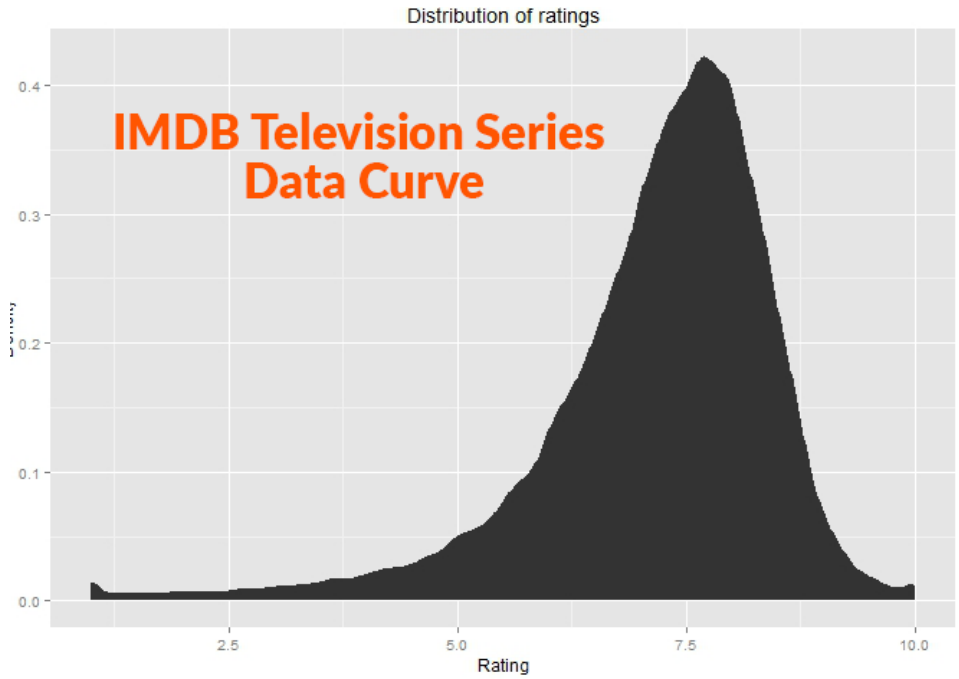
- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Amazon Studio

- Amazon Studio
  - The TV production company of Amazon
  - Want to use big data to find the great TV show

IMDB Rating Curve

- A rating curve of 2,500 TV shows on the website IMDB
- A television show with the rating of 9 points or above is considered a winner
- Amazon Studio want to make sure that it is on the right end of this curve

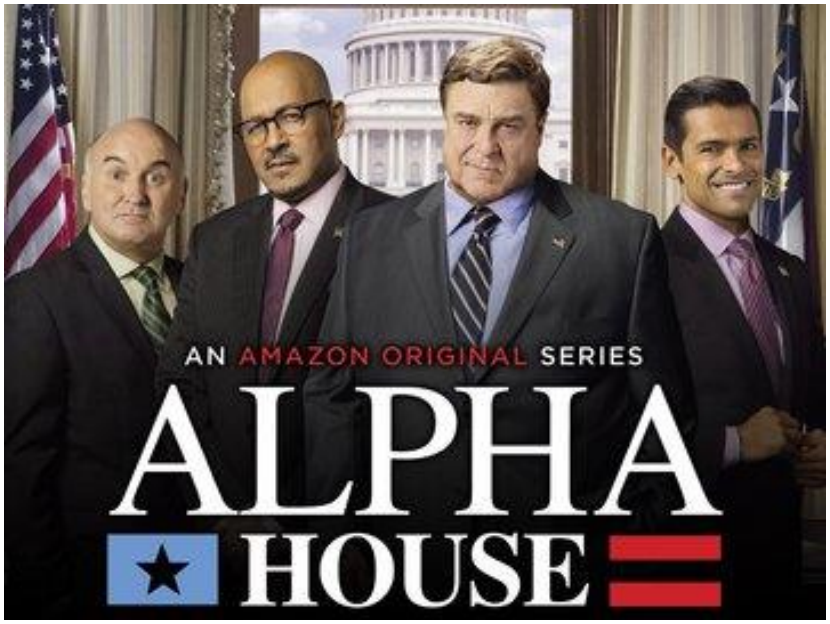


Alpha House

- Amazon Studio holds a competition and selects 8 candidates for TV shows
- Makes the first episode for each of these 8 shows and put them live on the internet for free
- Collects data of viewers behaviors and analyze the data to decide which show they should make and publicize

Alpha House

- Amazon decides to do a sitcom about four Republican US Senators, 'Alpha House'
- The show was not that good having a rating of 7.5, which was slightly above the average

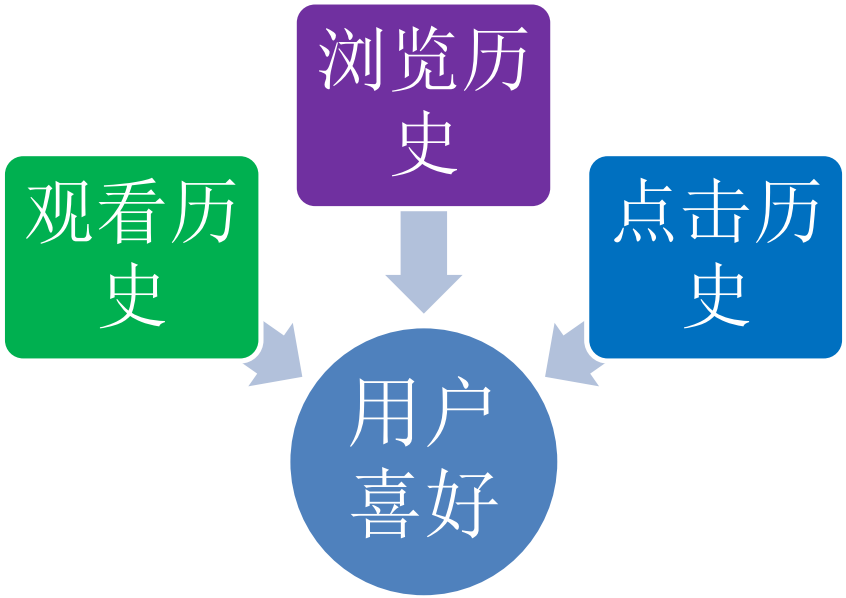


House of Cards

- Netflix went into looking at all the data they already had about their viewers
- Used that data to discover the bits and pieces about audience regarding what producers and actors they loved watching

House of Cards

- 数据分析发现喜欢经典英剧 House of Cards 的用户也很喜欢 Kevin Spacey 参演，或者 David Fincher 导演的作品。
- 决定投资翻拍有 Spacey 和 Fincher 参加的同名剧。



House of Cards

- Netflix brought all the pieces together and creat the very successful a drama series, “House of Cards.”
- Got a 9.1 rating on the IMDB Television Series Data Curve.

Differences Between Amazon and Netflix

- The difference between the approaches of these two most competitive data-savvy companies
- Netflix used data and brains together to make the decision of creating ‘House of Cards’
- Amazon used all those data to drive their decision making

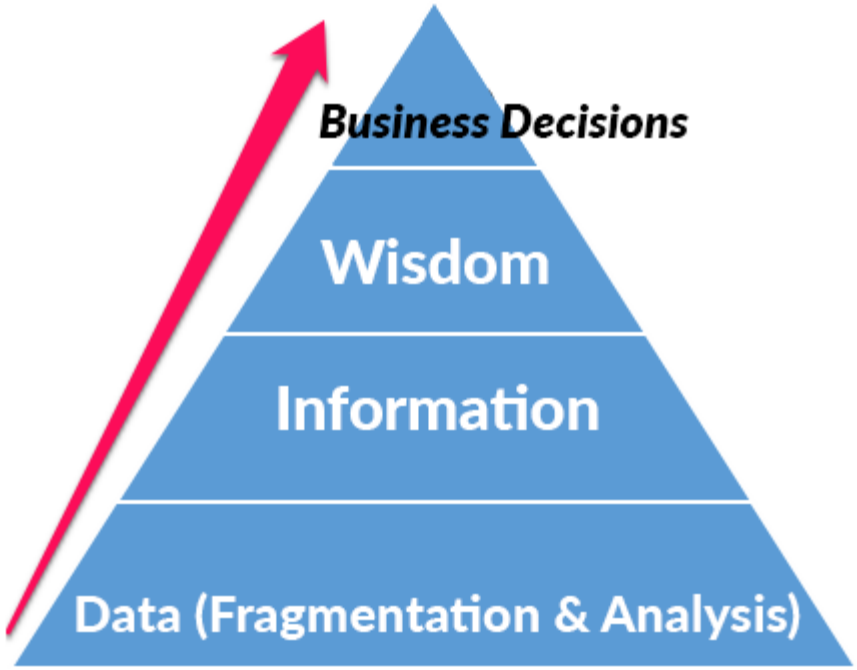
Data is just a tool

- Data and data analysis is only helpful in taking a problem apart and understanding its bits and pieces.
- It's the brain of a data expert that can put those bits and pieces together and deriving a great conclusion
- Data is certainly a massively useful tool for better decision making, but things go in the wrong direction when this data starts to drive those decisions

Data is just a tool

- The data is just the tool which help us to know better but not the key to success

Data Analysis+ Human Intelligence==Better Business Decisions



谢 谢

http://www.inpluslab.com

移动互联网与金融大数据实验室

数据挖掘里程碑

- 1763 年，Thomas Bayes 的论文在他死后发表
  - Bayes 理论将当前概率与先验概率联系起来
  - Bayes 理论能够帮助理解基于概率估计的复杂现况
  - 成为数据挖掘和概率论的基础
- 1805 年, Adrien-Marie Legendre 和 Carl Friedrich Gauss 使用回归确定了天体（彗星和行星）绕行太阳的轨道
  - 回归分析的目标是估计变量之间的关系
  - 在这个例子中采用的方法是最小二乘法
  - 回归成为数据挖掘的重要工具之一

数据挖掘里程碑

- 1936 年，计算机时代即将到来，海量数据的收集和处理成为可能
  - 1936年发表的论文《On Computable Numbers》中，Alan Turing 介绍了通用图灵机的构想
  - 通用机具有像今天的计算机一般的计算能力
  - 现代计算机就是在图灵这一开创性概念上建立起来的
- 1943 年，Warren McCullon 和 Walter Pitts 首先构建出神经网络的概念模型
  - 《A logical calculus of the ideas immanent in nervous activity》 的论文阐述了网络中神经元的概念
  - 每一个神经元可以做三件事情：接受输入，处理输入和生成输出。

数据挖掘里程碑

- 1975 年，John Henry Holland 所著的《自然与人工系统中的适应》问世
  - 成为遗传算法领域具有开创意义的著作
  - 讲解了遗传算法领域中的基本知识，阐述理论基础，探索其应用
- 1989 年，术语“数据库中的知识发现”（KDD）被Gregory Piatetsky-Shapiro 提出
  - 合作建立起第一个同样名为KDD的研讨会

数据挖掘里程碑

- 1992 年， Berhard E. Boser, Isabelle M. Guyon 和 Vladimir N. Vanik对原始的支持向量机提出了一种改进办法
  - 新的支持向量机充分考虑到非线性分类器的构建
- 1993 年， Gregory Piatetsky-Shapiro 创立“ Knowledge Discovery Nuggets (KDnuggets) ”通讯
  - 本意是联系参加KDD研讨会的研究者
  - KDnuggets.com的读者群现在似乎广泛得多

数据挖掘里程碑

- 2003 年，Micheal Lewis 写的 《点球成金》 出版
  - 奥克兰运动家队（美国职业棒球大联盟球队）使用一种统计的，数据驱动的方式针对球员的素质进行筛选，这些球员被低估或者身价更低
  - 成功组建了一支打进2002和2003年季后赛的队伍，而他们的薪金总额只有对手的1/3