

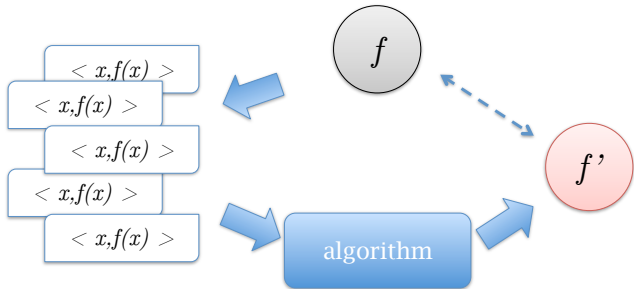


Lecture 1: Introduction

http://cs.nju.edu.cn/yuy/course_dm12.ashx



An abstract view



Data mining



“Data mining is the analysis of (often **large**) **observational** data sets to find **unsuspected relationships** and to summarize the data in **novel** ways that are both **understandable** and **useful** to the data owner.”

[D. Hand et al. , Principles of Data Mining]

数据挖掘是通过对(大规模)观测数据集的分析,寻找确信的关系,并将数据以一种可理解的且利于使用的新颖方式概括数据的方法。

Data mining factors



Large: small data needs no data mining

Unsuspected relationships: correct and significant

Novel: rediscovery of known facts is useless

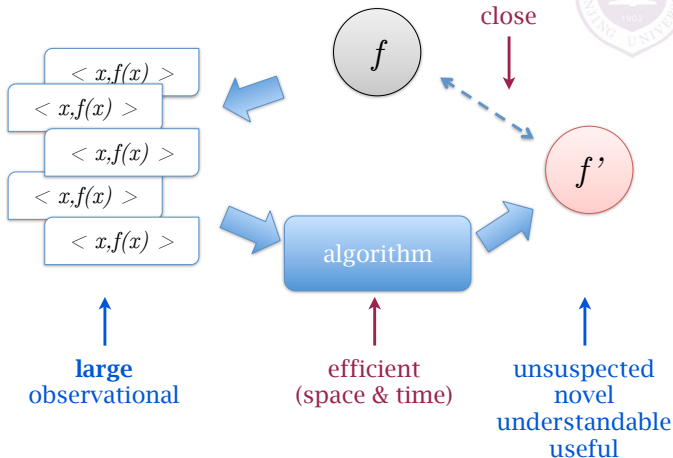
Understandable: decision maker oriented

Useful: mining results should be useful to the users

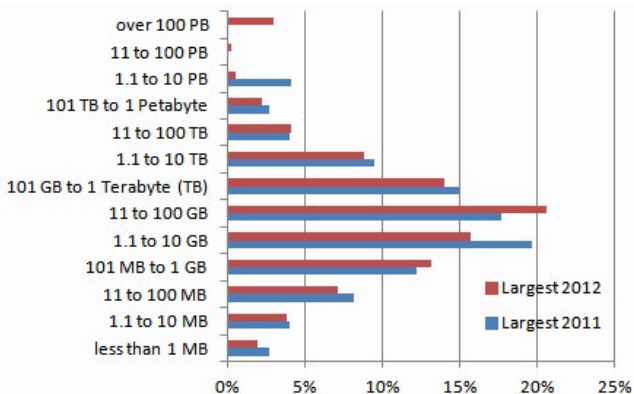
Observational data v.s. experimental data

[D. Hand et al. , Principles of Data Mining]

Data mining factors



How large can the data be



[KDnuggets Poll, 2012]

What can data mining do? DM Tasks



Exploratory data analysis

interactive and visualized

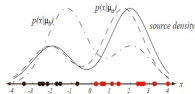
how to visualize high dimensional data?



Descriptive modeling

describe a data set

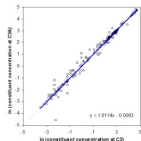
how to characterize general properties of a dataset



Predictive Modeling

perform inference from a data set

how to construct the mapping from the input space to the output space



Discovering patterns and rules

find association relationship

how to find high correlated items out of a huge data set

Retrieval by content



Example: Mining supermarket transactions



Example: Mining valuable customers



GSM



CDMA

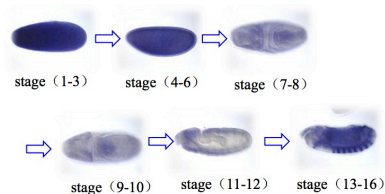


Example: mining network intrusion patterns



recognize intrusion accesses

Example: Mining biology data



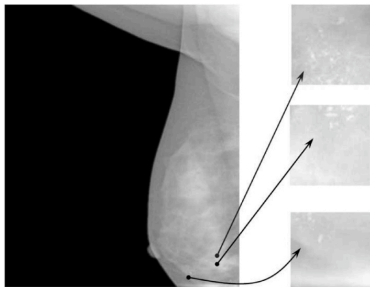
Finding key genes

Identifying gene expression patterns

Identifying gene interactions

...

Example: Mining medical data



Improving diagnosis of doctors by providing suggestions based on historical medical data

Example: Mining financial data



Fraud detection

Stock trends prediction

...

Example: Mining the web



Google™

bing™

amazon.com

Your Recent History (What's next?)

Recently Viewed Items

-  **Principles of Data Mining**
D. J. Hand
Hardcover
-  **Probabilistic Robotics**
Sebastian Thrun
Hardcover
-  **Simulation-based Algorithms for...**
Hongyi Soo Chang
Paperback
-  **Data Mining: Practical Machine...**
Ian H. Witten
Paperback

View and edit your browsing history

CeEnase Shopping: Customers Who Bought Items in Your Recent History Also Bought

Page 1 of 9



The Elements of Statistical Inference
+ Trevor Hastie
★★★★ (46)
Hardcover
\$63.05
[Fix this recommendation](#)



Machine Learning: An Algorithmic Approach
+ Stephen Marsland
★★★★ (21)
Hardcover
\$50.99
[Fix this recommendation](#)



Artificial Intelligence: A Modern Approach
+ Stuart J. Russell
★★★★ (41)
Hardcover
\$121.34
[Fix this recommendation](#)



Probabilistic Graphical Models
+ Daphne Koller
★★★★ (13)
Hardcover
\$82.00
[Fix this recommendation](#)



The Art of R Programming: A Tour...
+ Norman S. Matloff
★★★★ (22)
Paperback
\$24.32
[Fix this recommendation](#)



Pattern Classification (2nd Edition)
+ David G. Stork
★★★★ (33)
Hardcover
\$111.47
[Fix this recommendation](#)

Example: Mining usage data



Mining usage data to allow natural human-computer interaction

Top data mining fields



| Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters] | |
|--|------------------|
| | 2011 % of voters |
| CRM/ consumer analytics (57) | 25.0% |
| | 26.8% |
| Banking (43) | 18.9% |
| | 19.2% |
| Health care/ HR (38) | 16.7% |
| | 13.1% |
| Education (37) | 16.2% |
| | 9.9% |
| Fraud Detection (32) | 14.0% |
| | 12.7% |
| Science (31) | 13.6% |
| | 10.3% |
| Social Networks (30) | 13.2% |
| | 6.6% |
| Credit Scoring (29) | 12.7% |
| | 8.0% |
| Direct Marketing/ Fundraising (28) | 12.3% |
| | 11.3% |
| Insurance (28) | 12.3% |
| | 10.3% |
| Finance (26) | 11.4% |
| | 11.3% |
| Telecom / Cable (25) | 11.0% |
| | 10.8% |
| Retail (24) | 10.5% |
| | 8.0% |
| Medical/ Pharma (22) | 9.6% |
| | 8.0% |
| Biotech/Genomics (21) | 9.2% |
| | 5.6% |
| Government/Military (17) | 7.5% |
| | 6.1% |
| Travel / Hospitality (17) | 7.5% |
| | 1.4% |

Data types in mining tasks



“Flat” data: vectors and matrix

Show entries

Search:

| id | words | fog | kincaid | flesch | angel | animal | aristocracy | art | astronomy | beauty | being | cause | chance | change | citizen | constitution |
|-----------------------------|---------|-----|---------|--------|-------|--------|-------------|-----|-----------|--------|-------|-------|--------|--------|---------|--------------|
| aeschylus-agamemnon-1860 | 14951 | 8 | 6 | 80 | 0 | 0 | 0 | 3 | 0 | 0 | 118 | 1 | 2 | 0 | 0 | 0 |
| aeschylus-persians-1782 | 8372 | 14 | 11 | 62 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| aeschylus-prometheus-2549 | 10070 | 10 | 8 | 68 | 0 | 0 | 0 | 19 | 0 | 0 | 194 | 1 | 1 | 0 | 0 | 0 |
| aeschylus-seven-2836 | 9160 | 11 | 8 | 72 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 4 | 1 | 7 | 0 |
| aeschylus-suppliant-2642 | 9339 | 10 | 8 | 71 | 0 | 0 | 0 | 2 | 0 | 0 | 95 | 2 | 7 | 1 | 7 | 0 |
| american-articles-3758 | 3424 | 40 | 36 | -17 | 0 | 0 | 0 | 0 | 0 | 0 | 509 | 3 | 0 | 0 | 0 | 0 |
| american-constitution-4487 | 4517 | 22 | 19 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 535 | 0 | 0 | 0 | 45 | 69 |
| american-declaration-3934 | 1337 | 23 | 19 | 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 15 |
| aquinas-summa-2292 | 2510121 | 14 | 11 | 55 | 47 | 11 | 0 | 1 | 0 | 1 | 290 | 6 | 0 | 2 | 0 | 1 |
| aristophanes-achamians-2166 | 12954 | 10 | 7 | 64 | 0 | 1 | 0 | 1 | 0 | 2 | 109 | 2 | 0 | 0 | 13 | 0 |

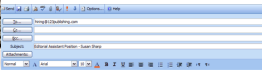
Showing 1 to 10 of 222 entries

First Previous 1 2 3 4 5 Next Last

Data types in mining tasks



Text data



Dear Hiring Manager,

I would like to express my interest in a position as editorial assistant for your publishing company. As a recent graduate with writing, editing, and administrative experience, I believe I am a strong candidate for a position at the 123 Publishing Company.

You specify that you are looking for someone with strong writing skills. As an English major, a writing tutor, and an editorial intern for both a government magazine and a college marketing office, I have become a skilled writer with a variety of experience.

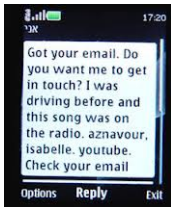
Although I am a recent college graduate, my maturity, practical experience, and eagerness to enter the publishing business will make me an excellent editorial assistant. I would love to begin my career with your company, and am confident that I would be a beneficial addition to the 123 Publishing Company.

I have attached my resume. Thank you so much for your time and consideration.

Sincerely,

Susan Sharp

Susan Sharp
123 Main Street
XYZ Town, NY 11111
Email: susan.sharp@gmail.com
Cell: 555-555-5555



Data types in mining tasks

Structured data



A screenshot of the CNN.com website. The main headline is "Israel hits Palestinian prime minister's office" with a photo of a man in a white shirt standing in a rubble-filled room. Other sections include "HIGHLIGHTS", "HIMALAYAN EXPRESS", and "INSIDE THE MIDDLE EAST". The CNN logo is prominent at the top left.

A screenshot of a Chinese website with a network diagram overlaid. The diagram consists of various colored nodes (purple, green, yellow, orange, red) connected by lines, representing a network structure. The website content includes a search bar and several categories of links, such as "高教系统", "南京大学", "多美教育", "网络教育", "学历教育", "文化艺术", "体育娱乐", "健康休闲", "新闻财经", and "综合广播".

Data types in mining tasks



Multi-media data



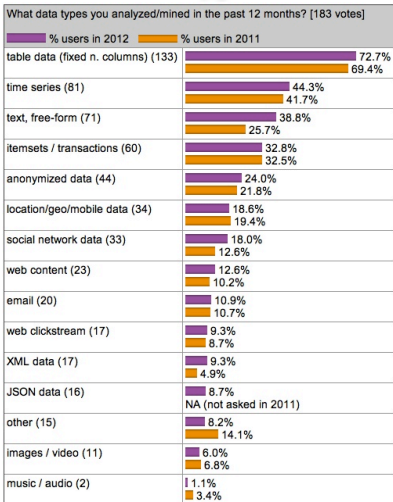
Data types in mining tasks



Temporal and spatial data



Top mined data types



Who needs data mining

Sponsors of ACM SIGKDD 2012 (Companies):



Nokia Research Center



SOSO 搜搜

PayPal of eBay



Baidu 百度

Deloitte.

CITYGRID.com



sas



阿里云
aliyun.com

Adobe

OPERA
SOLUTIONS

Microsoft

Google

technicolor

EMC²

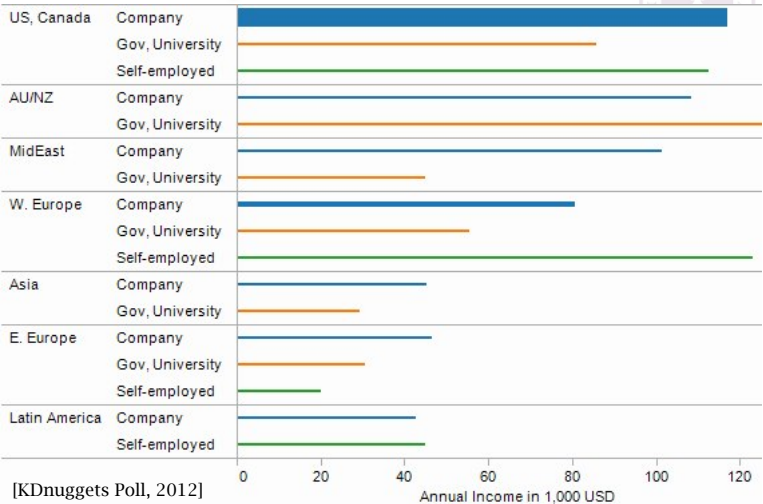
IBM Research

Greenplum

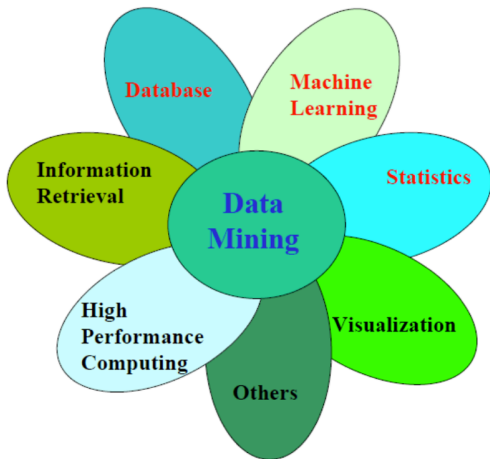
m6d
media6degrees

SALFORD
SYSTEMS

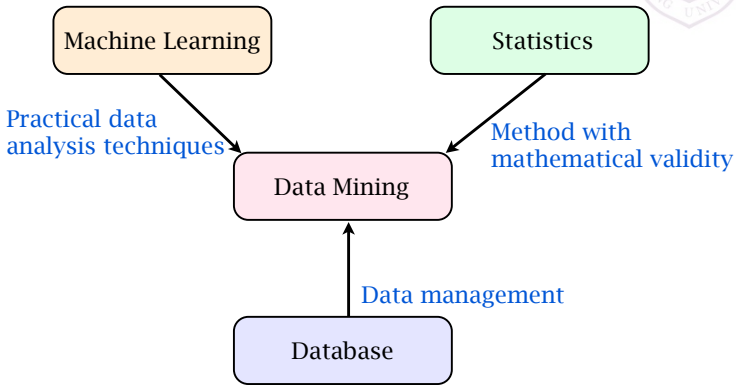
Annual salary of data miners



Cross-disciplines of data mining



Three perspectives of data mining



习题



为何数据挖掘强调挖掘大数据集？

为何强调数据挖掘结果的可理解性？

数据挖掘是否只处理表格数据？

数据挖掘与统计有哪些区别？