

C S 488/508: Introduction to Data Mining

Huiping Cao

Dept. of Computer Science

New Mexico State University

Instructor information

- Dr. Huiping Cao
- Email: hcao@nmsu.edu
- Webpage: <http://www.cs.nmsu.edu/~hcao>
- Office: Science Hall 171
- Tel: 575-646-4600 (office)
- Personal Zoom Meeting link: <https://nmsu.zoom.us/j/8549600124>

Why this course?

According to a McKinsey report¹, McKinsey predicts a great effect of big data in employment, where 140,000 – 190,000 workers with “deep analytical” experience will be needed in the U.S.; furthermore, 1.5 million managers will need to become data-literate.

1. M. Manyika et al. Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. May 2011.

8 Most In-Demand Jobs Every Company Will Be Hiring For in 2021

1. Data analysts will be in demand.

According to the report, data analysts will become increasingly more important in all industries by 2020.

Survey respondents said they expect to have a greater demand for data analysts because they will need help making sense of all of the data generated by technological disruptions.

2. In fact, computer and mathematical jobs as a whole will also continue to get a boost.

Jobs that fall under the computer and mathematical occupations will grow.

These occupations include computer programmers, software developers, information security analysts, and more.

Jun 11, 2021, 09:44am EDT | 5,237 views

The Data Analytics Profession And Employment Is Exploding—Three Trends That Matter

learning engineers. The U.S. Bureau of Labor Statistics sees strong growth for data science jobs skills in its prediction that the data science field will grow about 28% through 2026. Also, as technology improves, companies have been able to increase the sophistication of their data operations and analysis. Increasingly, that means inserting artificial intelligence (AI) capabilities into the business processes of regular companies (i.e. non-tech giants). And that means demand for data scientists (average salary in USA \$111,100) and related positions (research scientists and machine learning engineer) will also

8 of the Most In-Demand Engineering Jobs for 2021

By Dean McClements

01 June 2021

1. Data Science & Machine Learning

Software engineering has seen continuous growth over the past few years, and there are no signs of it stopping. Data science is a branch of software engineering that involves creating

So, without further ado, here are 2021's most in-demand engineering jobs and the salary potential one should expect from each.

1. Data Science & Machine Learning. ...
2. Automation & Robotics Engineer. ...
3. Mechanical Engineer. ...
4. Civil Engineer. ...
5. Electrical Engineer. ...
6. Alternative Energy Engineer. ...
7. Mining Engineer. ...
8. Project Engineer.

Course description & topic list

- Course description
 - Techniques for exploring large data sets and discovering patterns in them. Data mining **concepts, metrics** to measure its effectiveness. Methods in **classification, clustering, frequent pattern analysis**. Selected topics from current advances in data mining.
- Course topic list
 - (a) Data, data pre-processing, proximity
 - (b) Regression: linear
 - (c) Classification
 - (d) Clustering
 - (e) Association analysis
 - (f) Anomaly detection

Course objectives

- Upon completion of this course, students will:
 - LO1: Explain and recognize different **data mining tasks** such as data pre-processing, visualization, classification, regression, clustering, association rules, and anomaly detection
 - LO2: **Apply** classical data-mining/machine-learning algorithms for classification, clustering, association rules, and anomaly detection
 - LO3: **Evaluate and compare** the performance of different data-mining/machine-learning algorithms
 - LO4: **Utilize** data mining algorithms to analyze data in real applications using a **data mining tool**

Syllabus & Canvas

- Go through Syllabus
- Go through Canvas course

Outline

1. Motivations
2. What is data mining
3. Origins of data mining
4. Typical data mining tasks
5. Challenges

Motivation

- Vast amount
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all
- Data growth¹
 - Worldwide information is more than doubling every two years
 - 1.8 zettabytes or 1.8 trillion gigabytes projected to be created and replicated in 2011

1. Courtesy of Zdnet

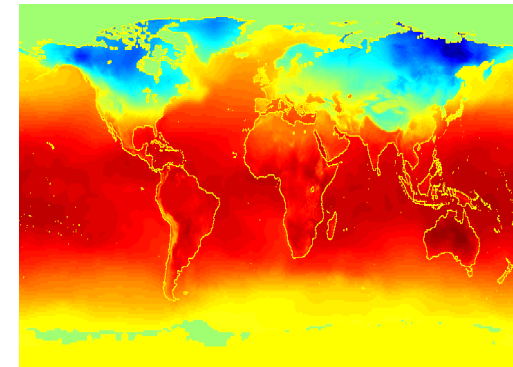
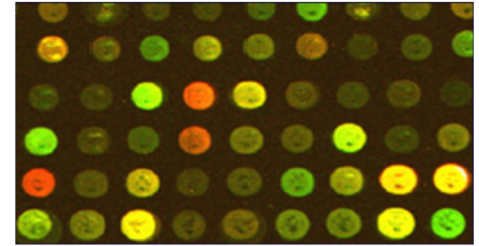
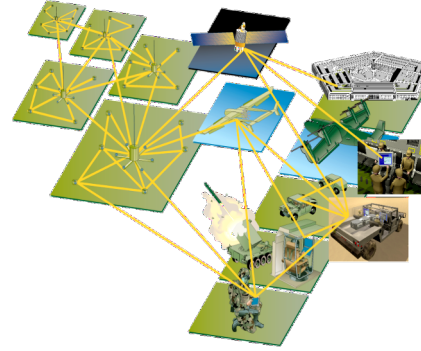
<http://www.zdnetasia.com/data-volume-to-hit-1-8zb-in-2011-62301103.htm>

Applications: Business

- Product Promotion
 - <https://www.kaggle.com/c/springleaf-marketing-response>
 - Offer customers loans using direct mails
 - Focus on the customers who are likely to respond and be good candidates for their services
 - Using a large set of anonymized features, predict which customers will respond to a direct mail offer.
- Coupon Purchase Prediction
 - <https://www.kaggle.com/c/coupon-purchase-prediction>
 - Using past purchase and browsing behavior, predict which coupons a customer will buy in a given period of time. The resulting models will be used to improve a recommendation system, so they can make sure their customers don't miss out on their next favorite thing.

Applications - Science

- Data collected and stored at enormous speed
 - remote sensors on a satellite
 - telescopes scanning the skies
 - microarrays generating gene expression data
 - scientific simulations generating terabytes of data
- Data mining may help scientists
 - in classifying and segmenting data
 - in hypothesis formation

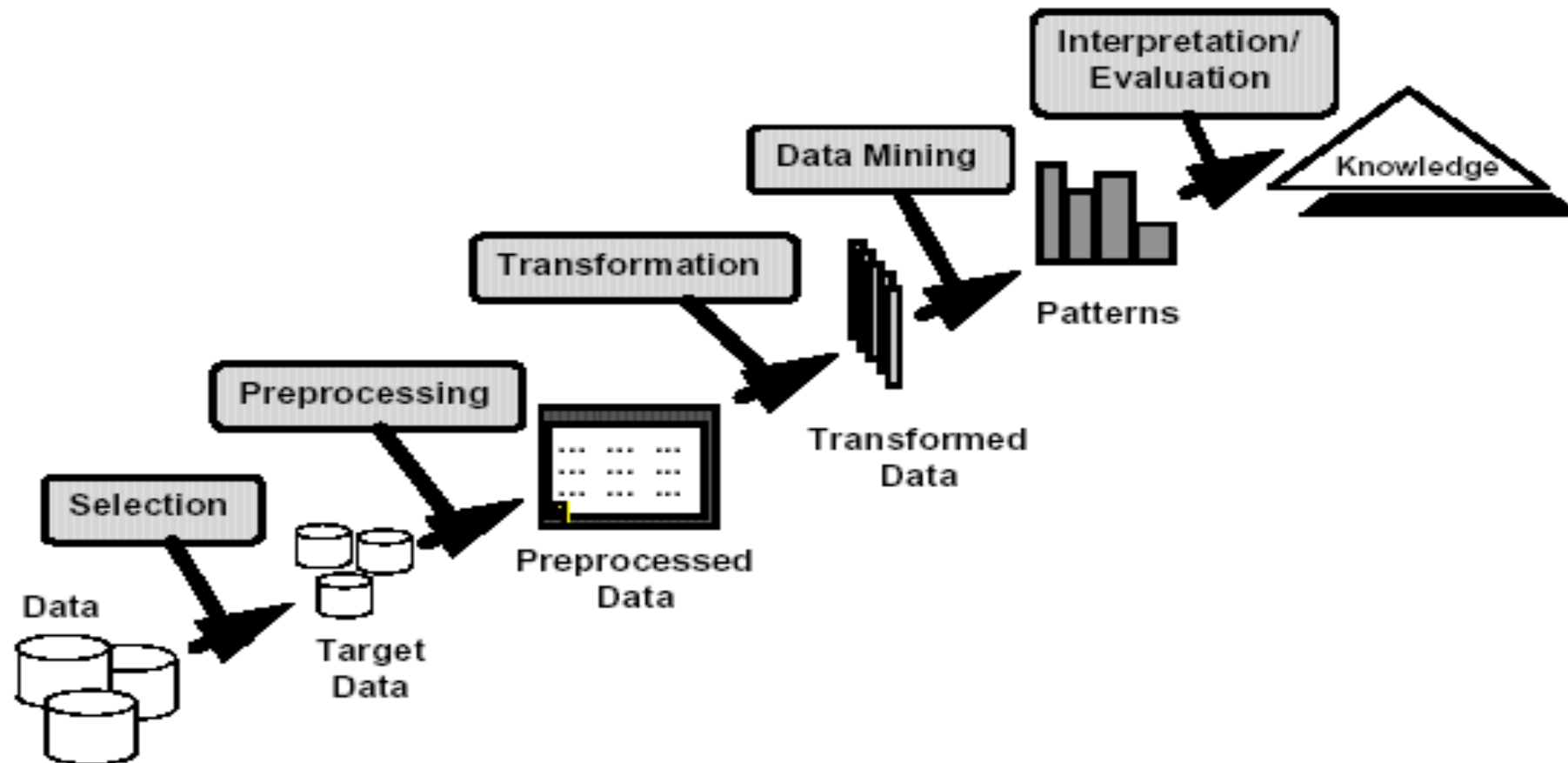


What is Data Mining - many definitions

- Non-trivial **extraction** of implicit, previously **unknown** and potentially useful information from data
- **Exploration & analysis**, by **automatic or semi-automatic** means, of large quantities of data in order to discover **meaningful** patterns

What is Data Mining (cont.)

- Knowledge discovery

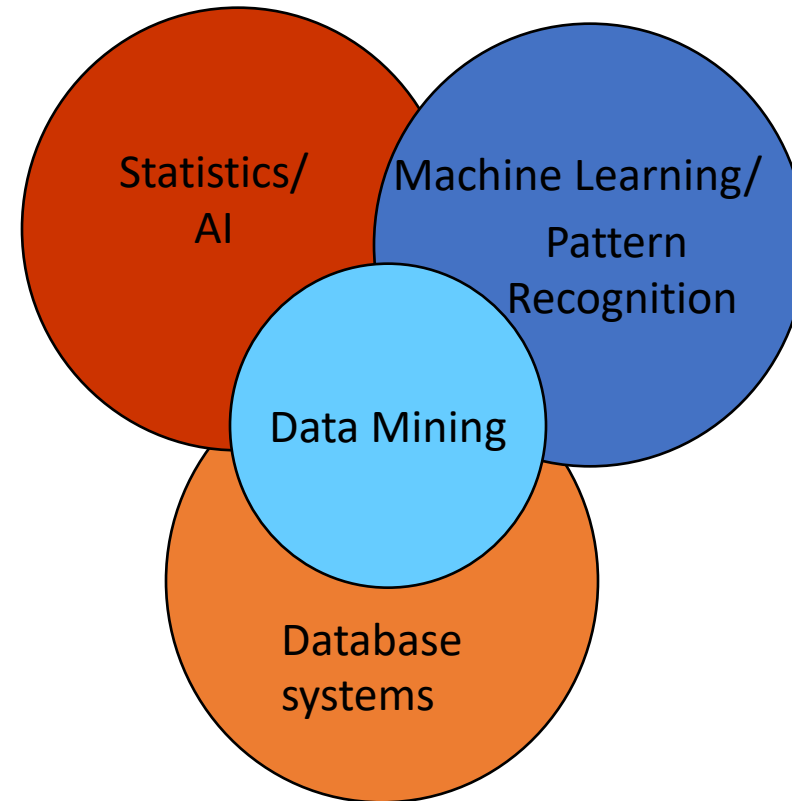


What is Data Mining (cont.)

- Discovering useful knowledge from large amount of data is data mining?
 - Not really
- Information retrieval
 - Look up phone number in phone directory
 - Query a Web search engine for information about “Amazon”

Origins of Data Mining

- Traditional Techniques may be unsuitable
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Typical Data Mining Tasks

- Classification [Predictive]
- Regression [Predictive]
- Cluster Analysis [Descriptive]
- Association Analysis [Descriptive]
- Anomaly Detection [Predictive]

Data Mining Tasks

- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.
 - Target, dependent variable
 - Explanatory, independent variable
- Description Methods
 - Find human-interpretable patterns that describe the data.

Classification

- 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 (test data)** should be assigned a class as accurately as possible.
 - A test set is used to determine the accuracy of the model.
- Examples
 - Credit card fraud detection
 - Disturbance classification
 - Animal category

Classification Example

<i>Tid</i>	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

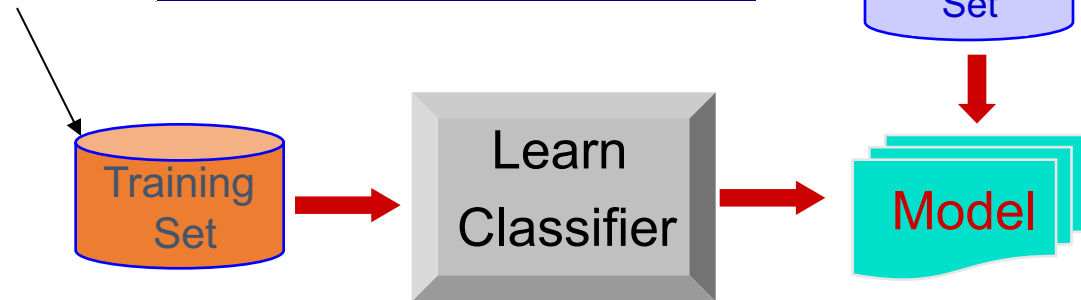
categorical

categorical

continuous

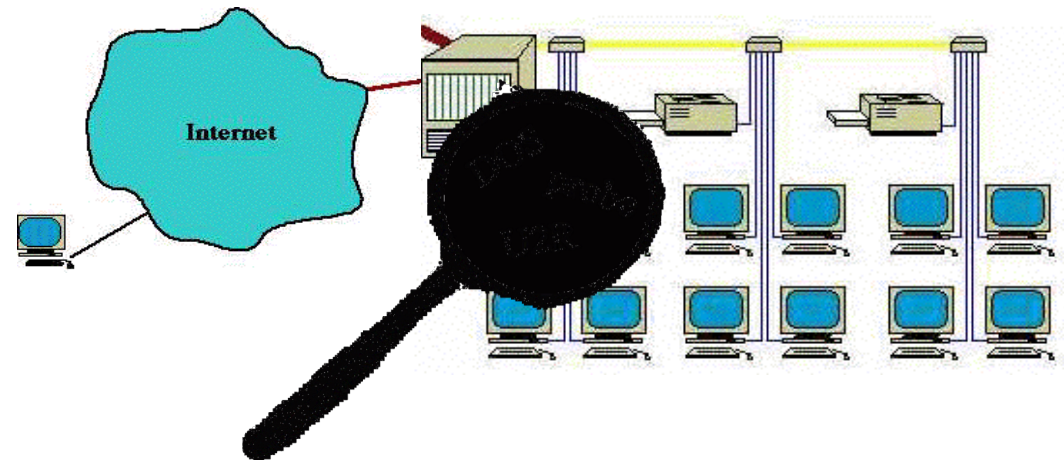
class

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



Deviation/Anomaly Detection

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



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.
 - Extensively studied in statistics, machine learning areas.
- Applications
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting power generation as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Regression Example

- Predict house price given 76 explanatory variables
- <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Bedroom number	Area	YearBuiltt	Price
3	856	2003	208500
3	1262	1976	181500
4	920	2002	223500
...

Bedroom number	Area	YearBuiltt	Price
2	900	2000	???
3	1000	1980	???

Cluster Analysis

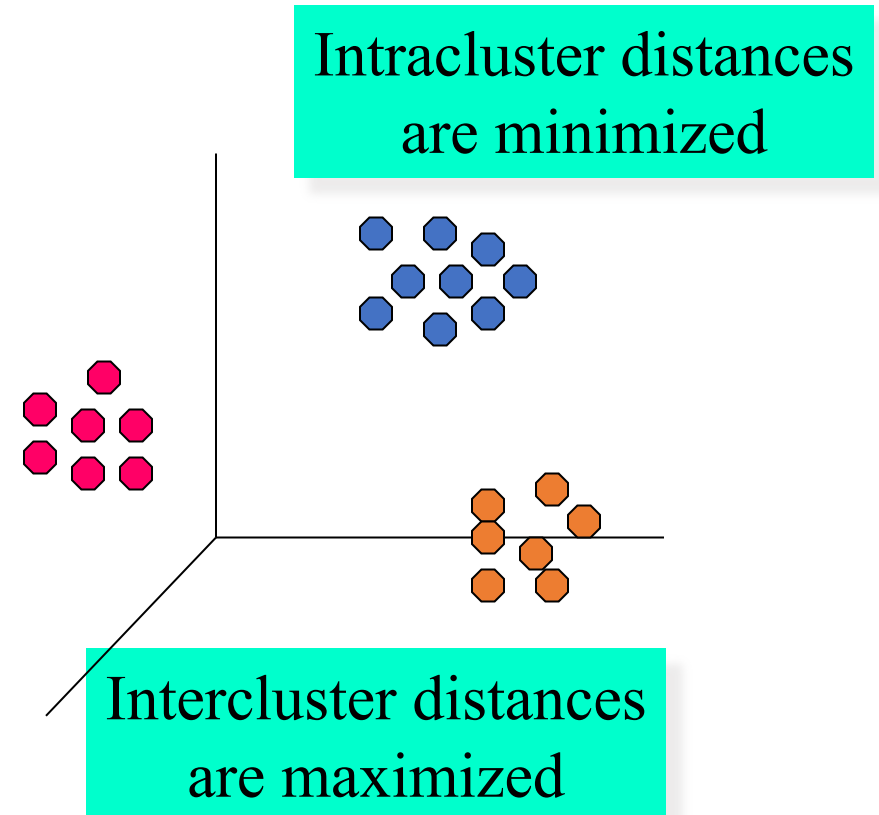
- What if I do not have training data?
- But I still want to categorize the data somehow...

Clustering: Application

- **Document clustering:** find **groups of documents** that are similar to each other based on the important terms appearing in them.
- **Animal behavior analysis:** group the animal movement data to different groups such that each group represents different behavior (e.g., resting ruminating, walking, drinking)

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.



Association Rule Discovery: Definition

- 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

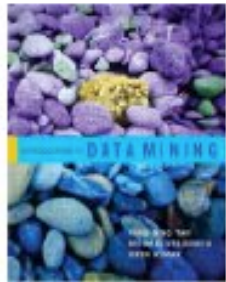
Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

Association rule discovery: Application

Frequently Bought Together



+



Price for both: **\$154.50**

Add both to Cart

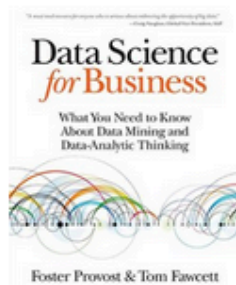
Add both to Wish List

One of these items ships sooner than the other. [Show details](#)

- ✓ **This item:** Introduction to Data Mining by Pang-Ning Tan Hardcover **\$118.91**
- ✓ Data Science for Business: What you need to know about data mining and data-analytic thinking by Foster Provost Paperback **\$35.59**

Customers Who Bought This Item Also Bought

Page 1 of 20



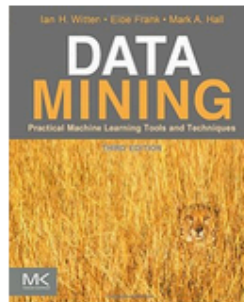
Data Science for Business:
What you need to know
about data mining and...

› Foster Provost

★★★★★ 114

Paperback

\$35.59 ✓Prime



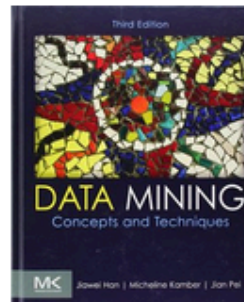
**Data Mining: Practical
Machine Learning Tools
and Techniques, Third...**

› Ian H. Witten

★★★★★ 59

Paperback

\$37.50 ✓Prime



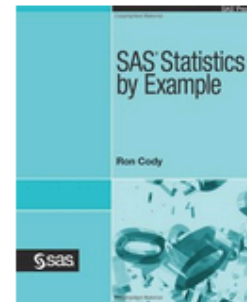
**Data Mining: Concepts and
Techniques, Third Edition
(The Morgan Kaufmann...**

› Jiawei Han

★★★★★ 29

Hardcover

\$50.02 ✓Prime

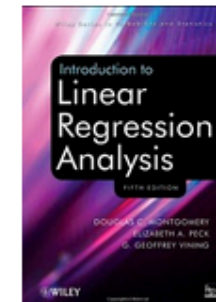


SAS Statistics by Example
Ron Cody

★★★★★ 11

Perfect Paperback

\$45.93 ✓Prime



**Introduction to Linear
Regression Analysis**

Douglas C. Montgomery

★★★★★ 17

Hardcover

\$93.08 ✓Prime



Types of data

- Relational tables (e.g., customers)
- Time series (e.g., voltage waveforms, speech)
- Images (e.g., CT scan)
- Text (e.g., documents)
- Graphs (e.g., social networks)
- Videos

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and heterogeneous Data
- Data distribution
- Communication, result consolidation, security
- Privacy preserving

References (1)

- Journals
 - TDKE, DMKD, DKE, KAIS
 - VLDBJ, IS (domain-specific)
- Conferences
 - Data mining: KDD, ICDM, SDM, PAKDD, PKDD
 - Artificial intelligence: IJCAI, AAI, UAI
 - Databases: SIGMOD, VLDB, ICDE, EDBT
 - Machine learning: NeurIPS, ICML
 - Information retrieval: SIGIR
 - Web related: WWW, WSDM
 - Natural language processing...
- Misc: CIKM

References (2)

- **Python:** <https://www.python.org/>
- **WEKA:** <http://www.cs.waikato.ac.nz/ml/weka/>
- **R:** <http://www.r-project.org/>
- **RapidMiner:** <http://rapidminer.com/products/rapidminer-studio/>
- **Apache Mahout:** a scalable machine learning library <https://mahout.apache.org/>

- **MATLAB:** <http://www.mathworks.com/>
- **SPSS:** <http://www-01.ibm.com/software/analytics/spss/>
- **SAS** (Statistical Analysis System): <http://www.sas.com/>
- **STATA** (Data Analysis and Statistical Software): <http://www.stata.com/>
- **Database** management systems

- **Kaggle** competition: <https://www.kaggle.com/competitions>