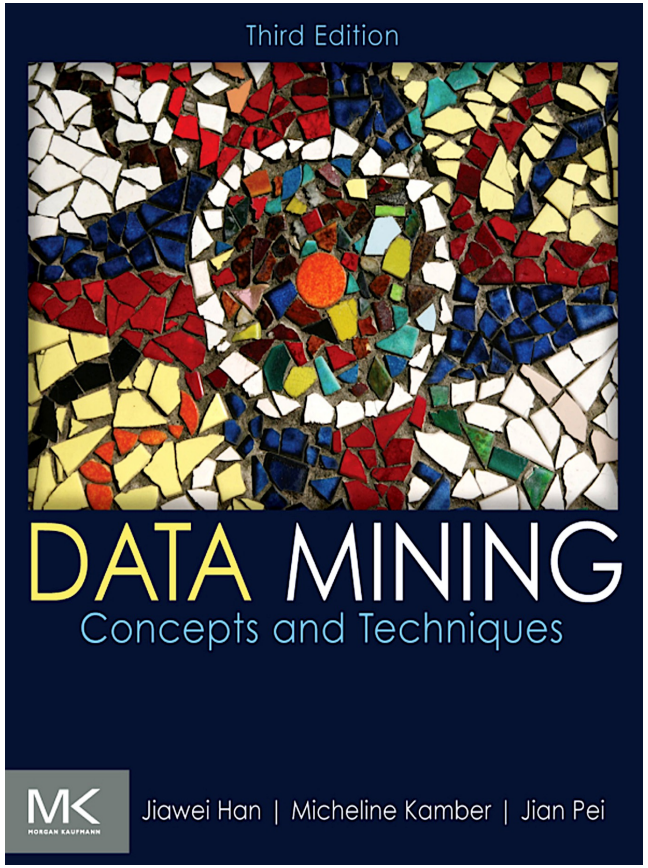
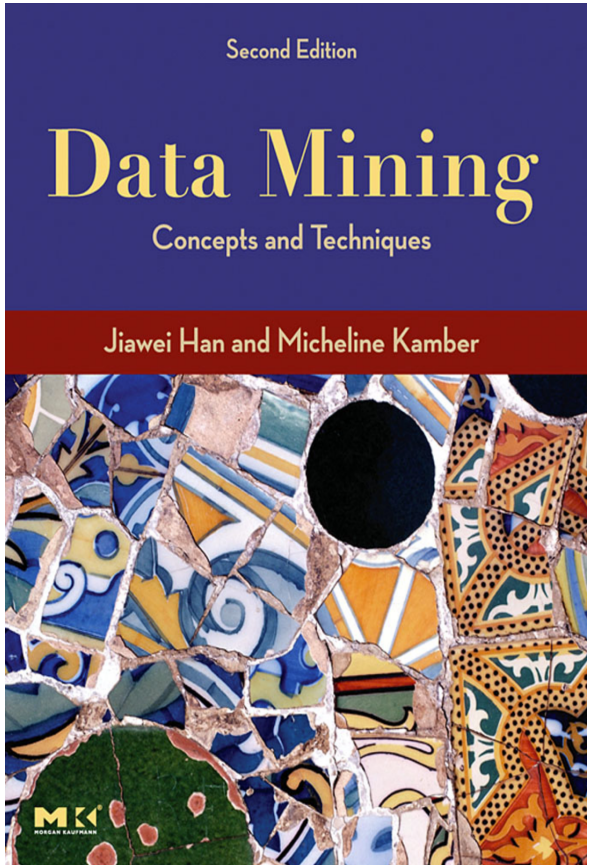
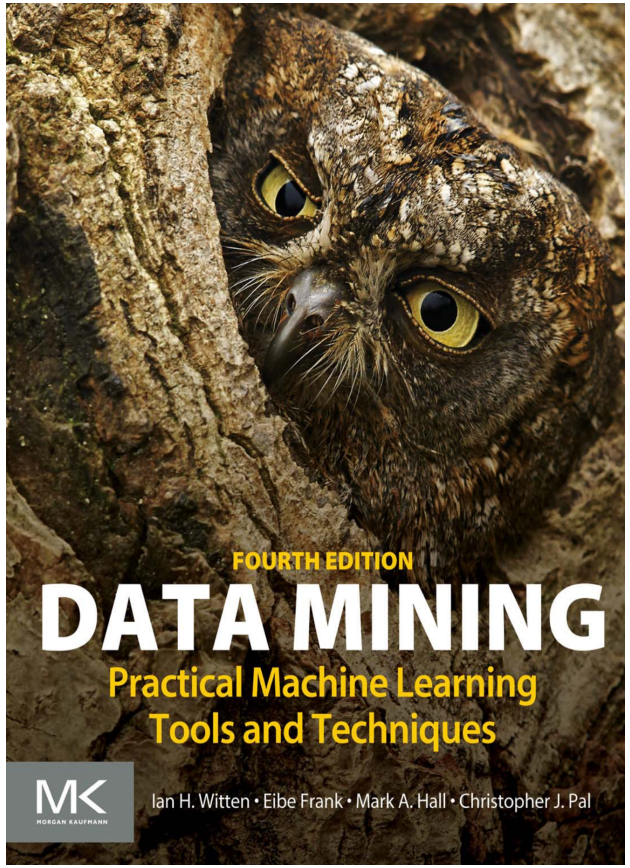
Data mining

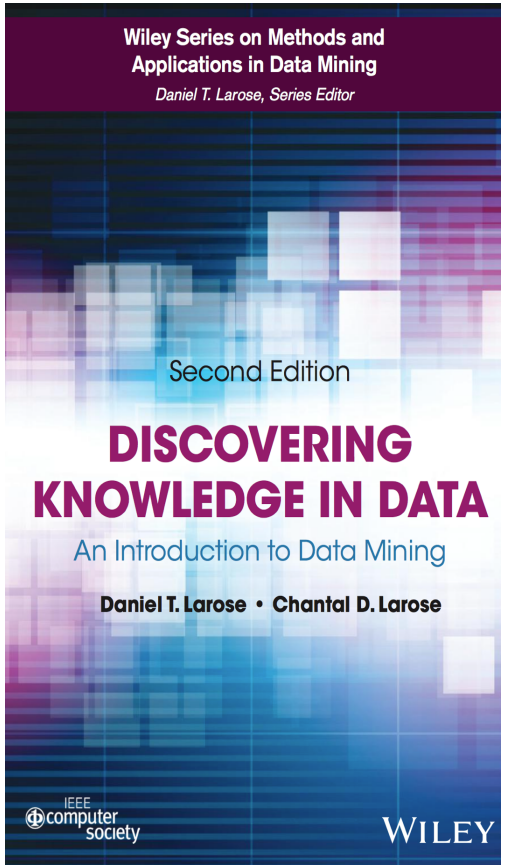
Introduction

References


Third Edition
DATA MINING
Concepts and Techniques
Jiawei Han | Micheline Kamber | Jian Pei


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber


FOURTH EDITION
DATA MINING
Practical Machine Learning
Tools and Techniques
Ian H. Witten • Eibe Frank • Mark A. Hall • Christopher J. Pal


Wiley Series on Methods and Applications in Data Mining
Daniel T. Larose, Series Editor
Second Edition
DISCOVERING KNOWLEDGE IN DATA
An Introduction to Data Mining
Daniel T. Larose • Chantal D. Larose
WILEY

Grading

- Presentation: 2 points
- Exercises: 5 points
- Final exam and quizzes: 13 points

Data

- The MIT Technology Review reports that it was the Obama campaign's effective use of data mining that helped President Obama win the 2012 presidential election.
  - They first identified likely Obama voters using a data mining model, and then made sure that these voters actually got to the polls.
  - They also used a separate data mining model to predict the polling outcomes. In the important swing county of Hamilton County, Ohio, the model predicted that Obama would receive 56.4% of the vote; the Obama share of the actual vote was 56.6%.
  - **This knowledge is used to allocate scarce resources more efficiently.**

Data

- The McKinsey Global Institute (MGI) reports (in 2011) that most American companies with more than 1000 employees had an average of at least 200 terabytes of stored data.

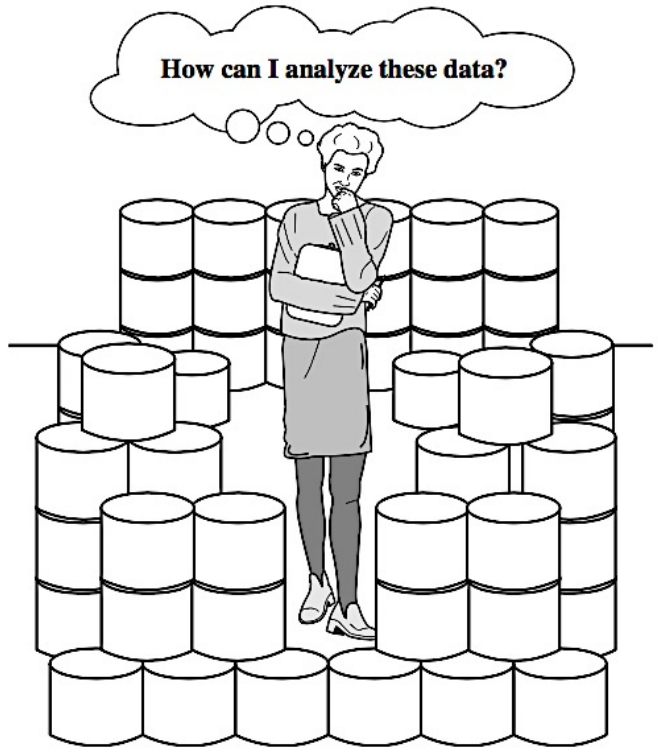
- MGI projects that the amount of data generated worldwide will increase by 40% annually.
  - Profitable opportunities for companies to leverage their data to reduce costs and increase their bottom line.

Data

- As early as 1984, in his book Megatrends, John Naisbitt observed that

"We are drowning in information but starved for knowledge."

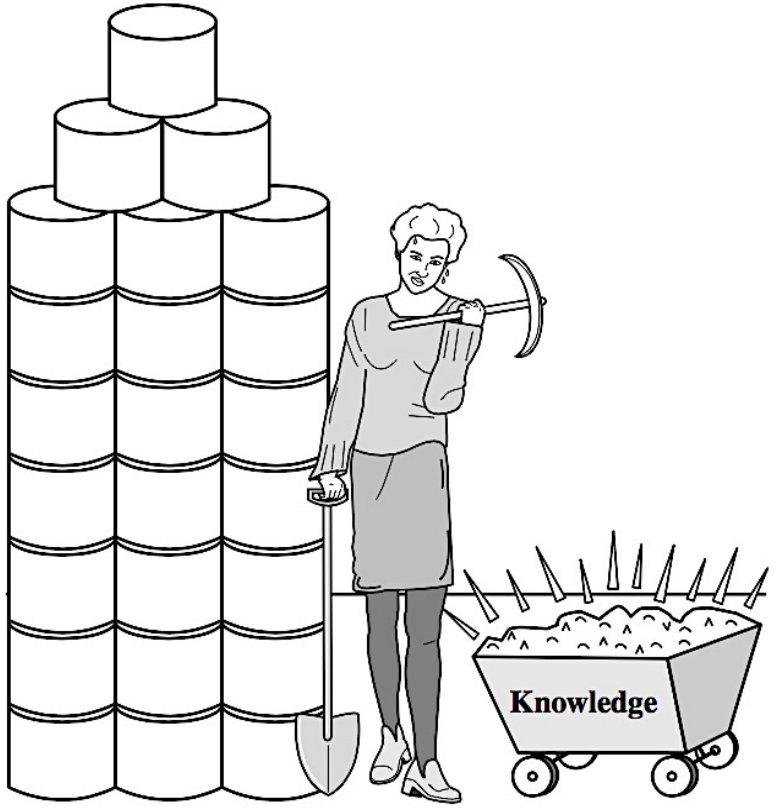
- There is a serious shortage of skilled data analysts.



The world is data rich but information poor.

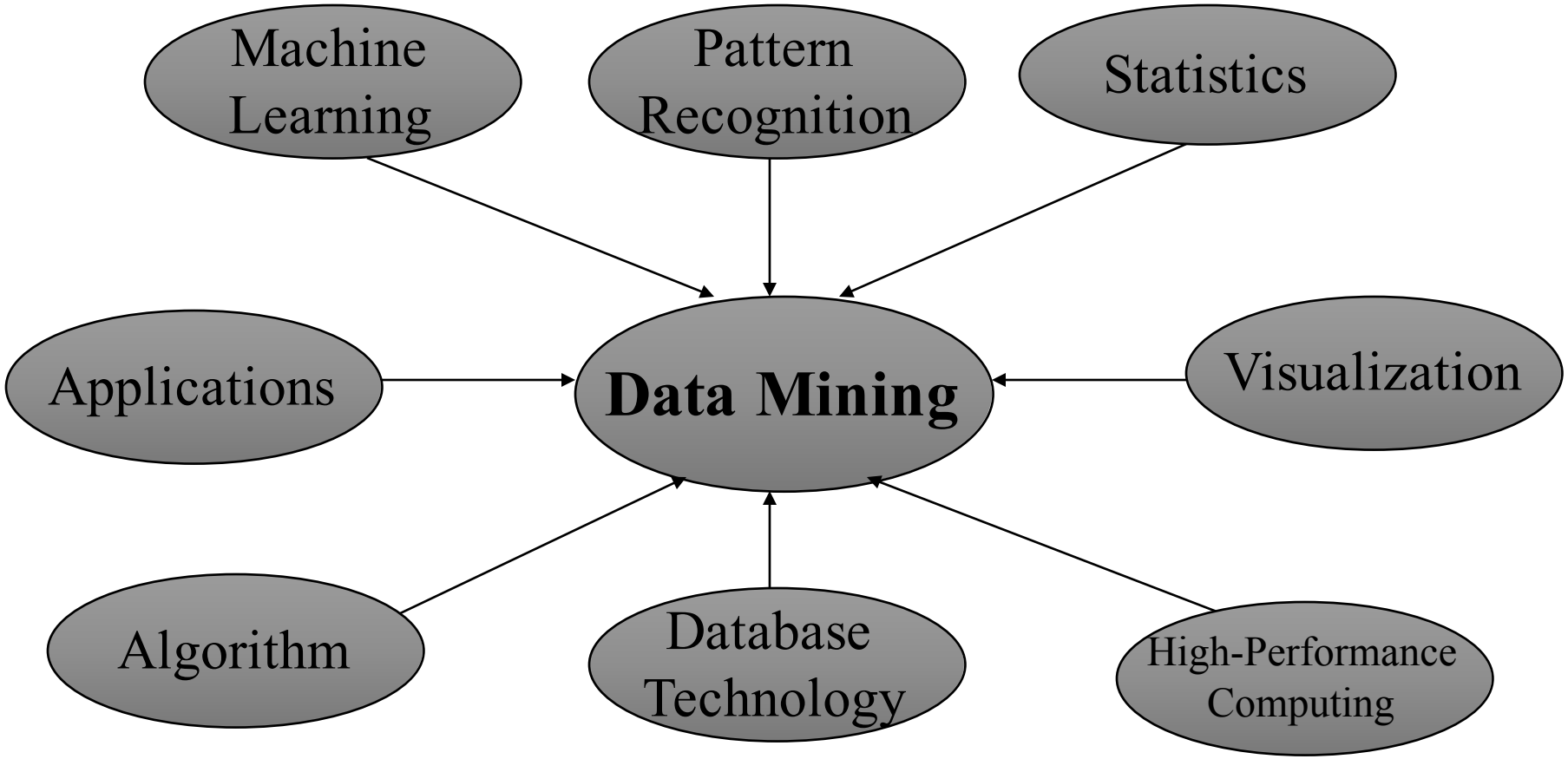
Data mining

Data mining is the process of discovering useful patterns and trends in large data sets.



Data mining—searching for knowledge (interesting patterns) in data.

Data mining: Confluence of Multiple Disciplines



Why now?

- The ongoing remarkable growth in the field of data mining and knowledge discovery has been fueled by a fortunate confluence of a variety of factors:
  - The explosive growth in data collection,
  - The storing of the data in data warehouses, so that the entire enterprise has access to a reliable, current database,
  - The availability of increased access to data from web navigation and intranets,
  - The competitive pressure to increase market share in a globalized economy,
  - The development of “off-the-shelf” commercial data mining software suites,
  - The tremendous growth in computing power and storage capacity.

Human role in data mining

- Berry and Linoff, in their 1997 book gave the following definition for data mining:
  - “Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules”.

- Three years later, in their Mastering Data Mining book, they mentioned that,
  - “If there is anything we regret, it is the phrase ‘by automatic or semi-automatic means’ . . . because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered.”

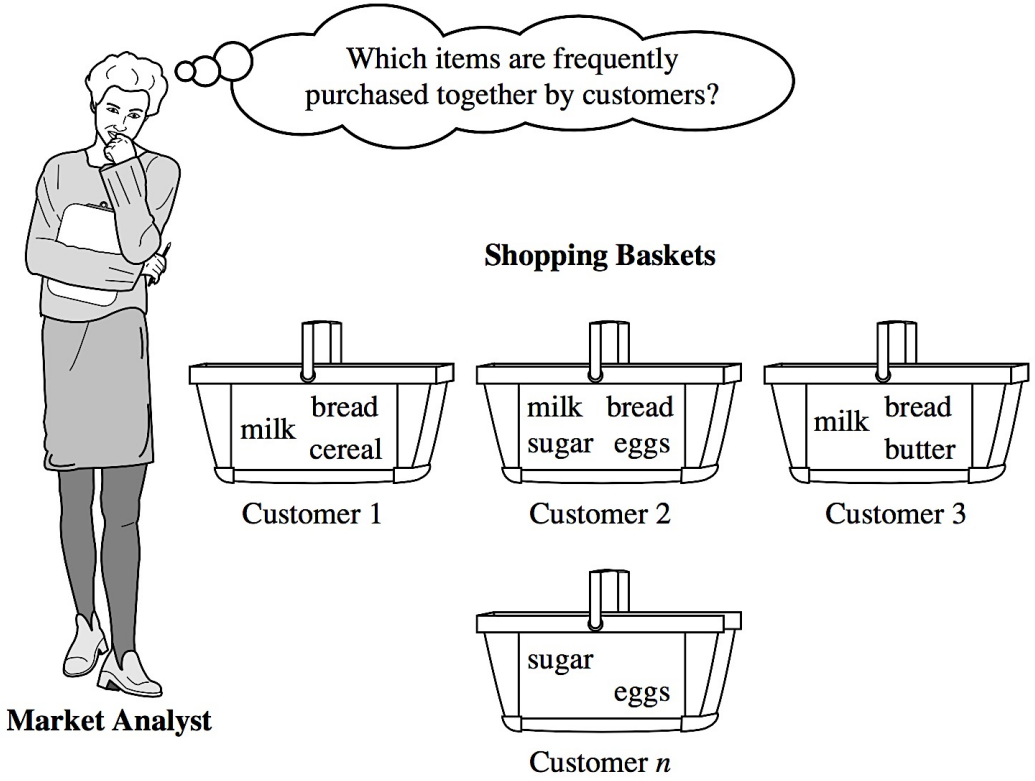
Human role in data mining

- Humans need to be actively involved at every phase of the data mining process.

- Models may be derived that are built upon wholly specious assumptions. Therefore, an understanding of the statistical and mathematical model structures underlying the software is required.

Examples

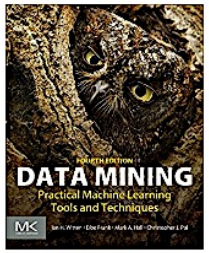
Market Basket Analysis



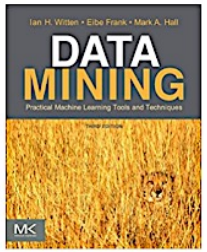
Examples

Customers who bought this item also bought

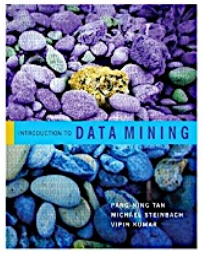
Page 1 of 20



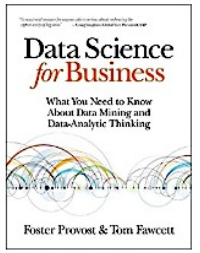
Data Mining, Fourth Edition: Practical Machine Learning Tools and...
› Ian H. Witten
5
Paperback
$33.31 prime



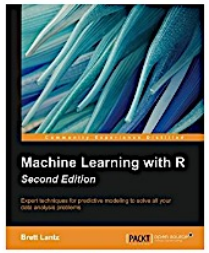
Data Mining: Practical Machine Learning Tools and Techniques, Third...
› Ian H. Witten
78
Paperback
36 offers from $28.60



Introduction to Data Mining
› Pang-Ning Tan
57
Hardcover
$150.95 prime

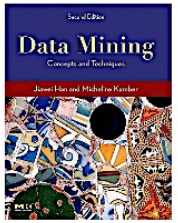


Data Science for Business: What You Need to Know about Data Mining and...
› Foster Provost
180
Paperback
$32.01 prime



Machine Learning with R - Second Edition: Expert techniques for predictive...
Brett Lantz
79
Paperback
$49.42 prime

goodreads Home My Books Browse Community Search books



Want to Read

Rate this book

Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)

by Jiawei Han, Micheline Kamber

3.85 · Rating details · 270 Ratings · 16 Reviews

Our ability to generate and collect data has been increasing rapidly. Not only are all of our business, scientific, and government transactions now computerized, but the widespread use of digital cameras, publication tools, and bar codes also generate data. On the collection side, scanned text and image platforms, satellite remote sensing systems, and the World Wide

Share

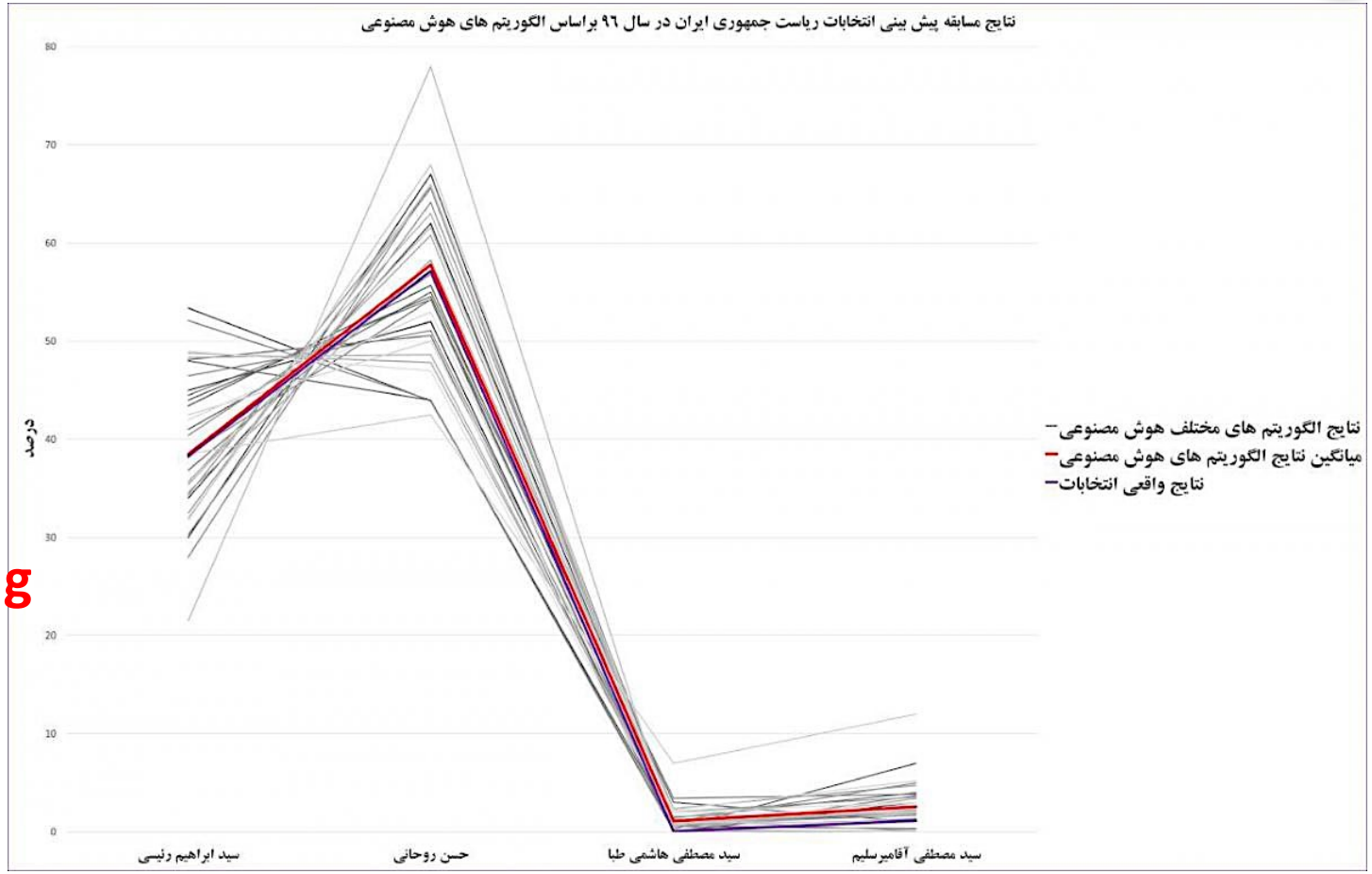
Recommend It | Stats | Recent Status Updates

READERS ALSO ENJOYED

Examples

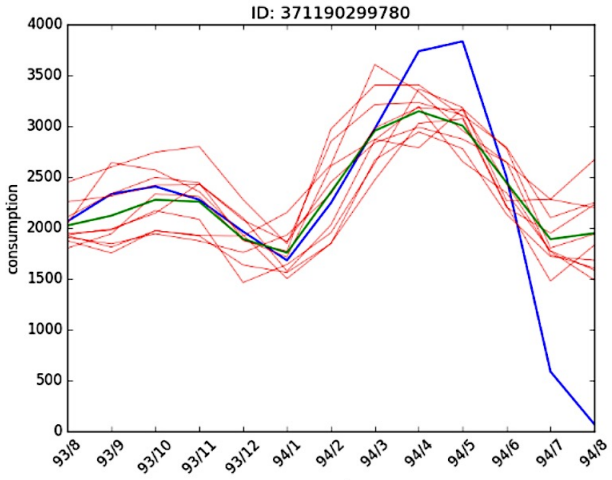
Predict the outcome of the election

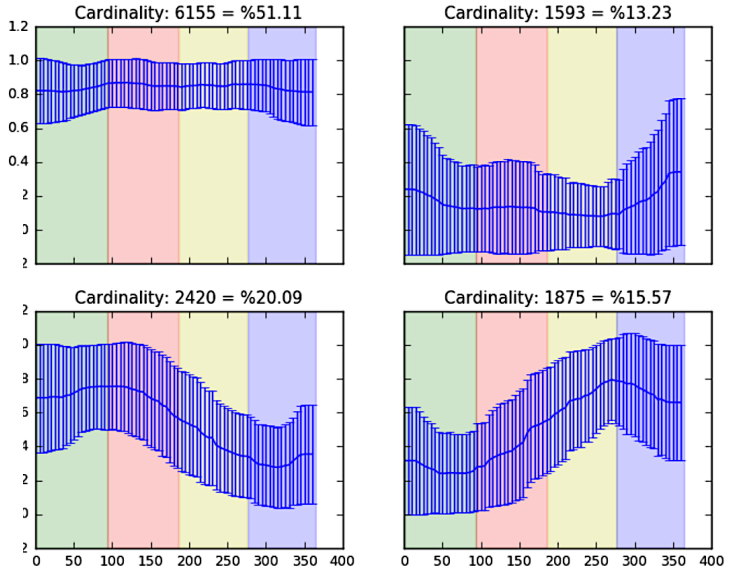
Example of **Opinion Mining** which is an example of **Text Mining**



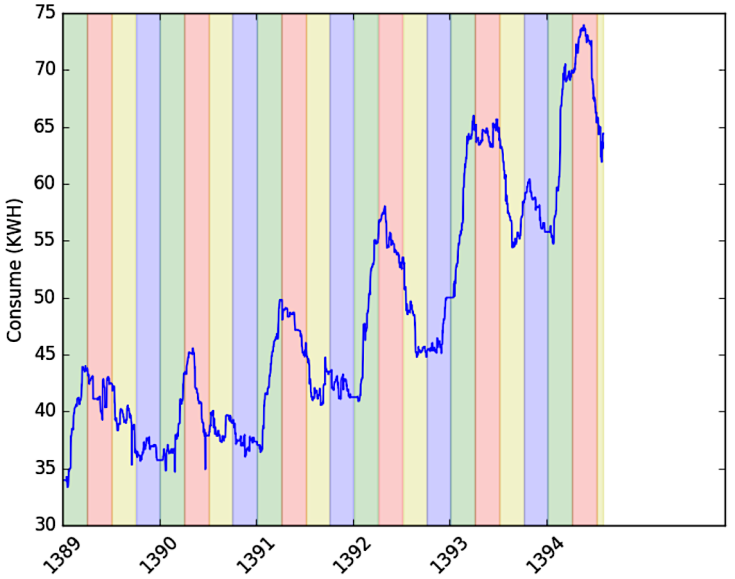
Examples

Bills Explorations





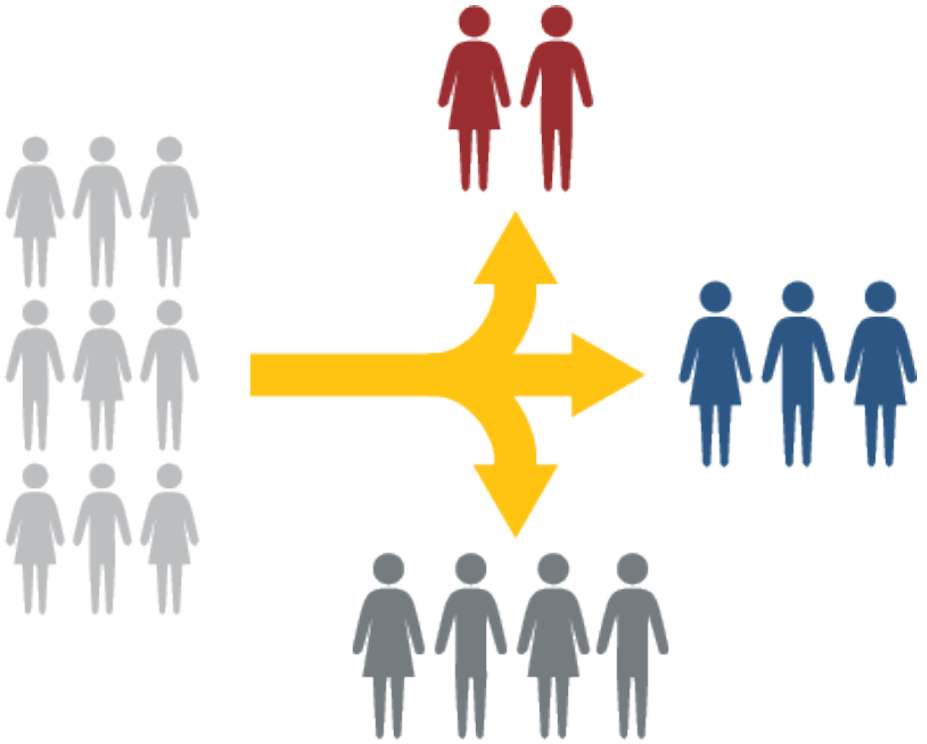
خوشه‌های مشترکین تک‌فاز تعرفه‌ی خانگی



نمودار میانگین مصرف مشترکین سه‌فاز تعرفه‌ی صنعت و معدن منطقه‌ی

Examples

Customer Segmentation



Examples

Sentiment Analysis

Example of **Text Mining**

مدل فول اچ دی:

از لحاظ داغ نکردن ونیومدن صدای فن عالیه و از لحاظ سرعت و صفحه نمایش معمولی ومتوسط واز لحاظ وزن مقداری سنگین میباشد.

Examples

CTR prediction (Click-through rate)

دکا چیست؟

امروزه پوش نوتیفیکیشن یکی از قوی‌ترین ابزارهای است که در دنیای اپلیکیشن‌ها و سرویس‌های موبایل برای جذب کاربران وجود دارد. اما همین ابزار قوی می‌تواند خطرآفرین هم بشود. ارسال اعلان به کاربری که علاقه‌ای به آن ندارد، او را آزار خواهد داد و احتمال حذف اپلیکیشن و از دست دادن کاربر را بالا خواهد برد.

در چنین موقعیتی، هوشمندی در ارسال و تشخیص مخاطبِ‌مناسب هر اعلان اهمیت ویژه‌ای پیدا می‌کند. در این مسابقه می‌خواهیم بر اساس داده‌های به‌دست‌آمده از تاریخچه‌ی کلیک کردن یا نکردن کاربران روی اعلان‌های مختلف، احتمال کلیک آن‌ها روی اعلان‌های جدید را پیش‌بینی کنیم.

Examples

Web Mining

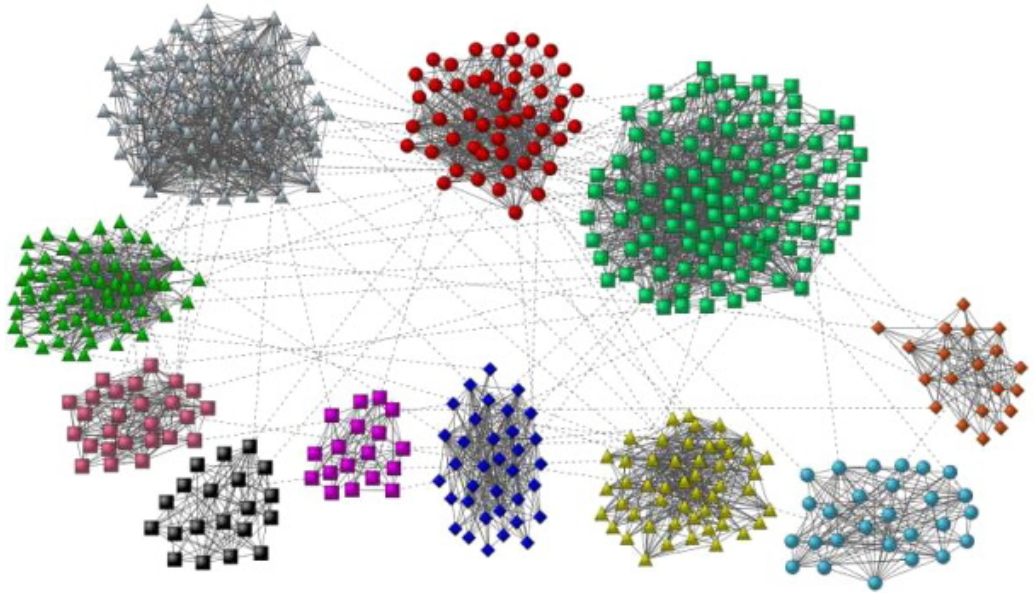
Example: J. Ginsberg, et al. “Detecting influenza epidemics using search engine query data,” *Nature*, 2009.

Examples

Credit Card
Fraud Detection

Examples

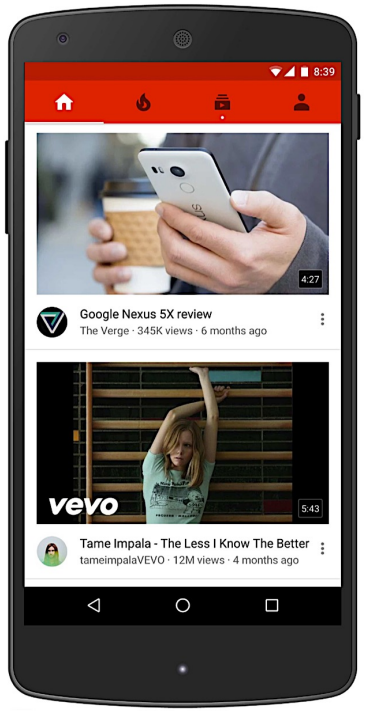
Graph Mining

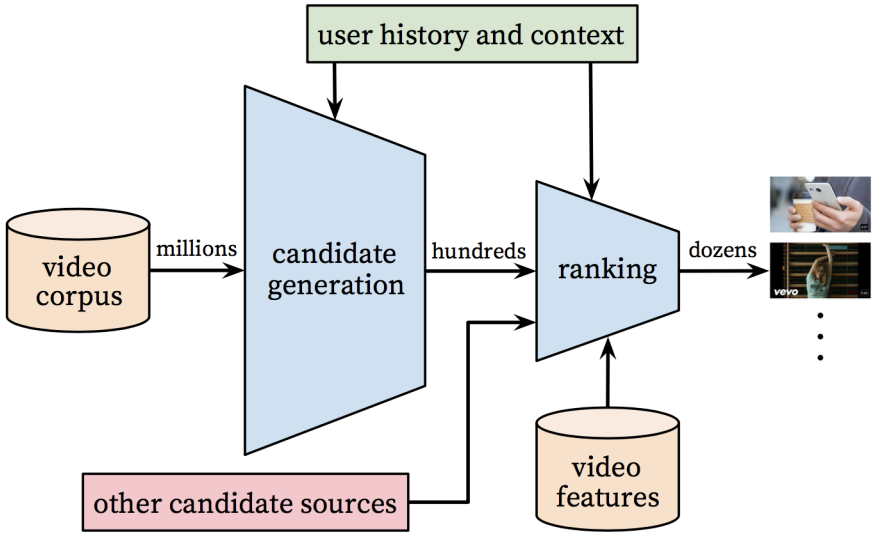


Useful in Anti Money Laundering Systems.

Examples

Recommender Systems



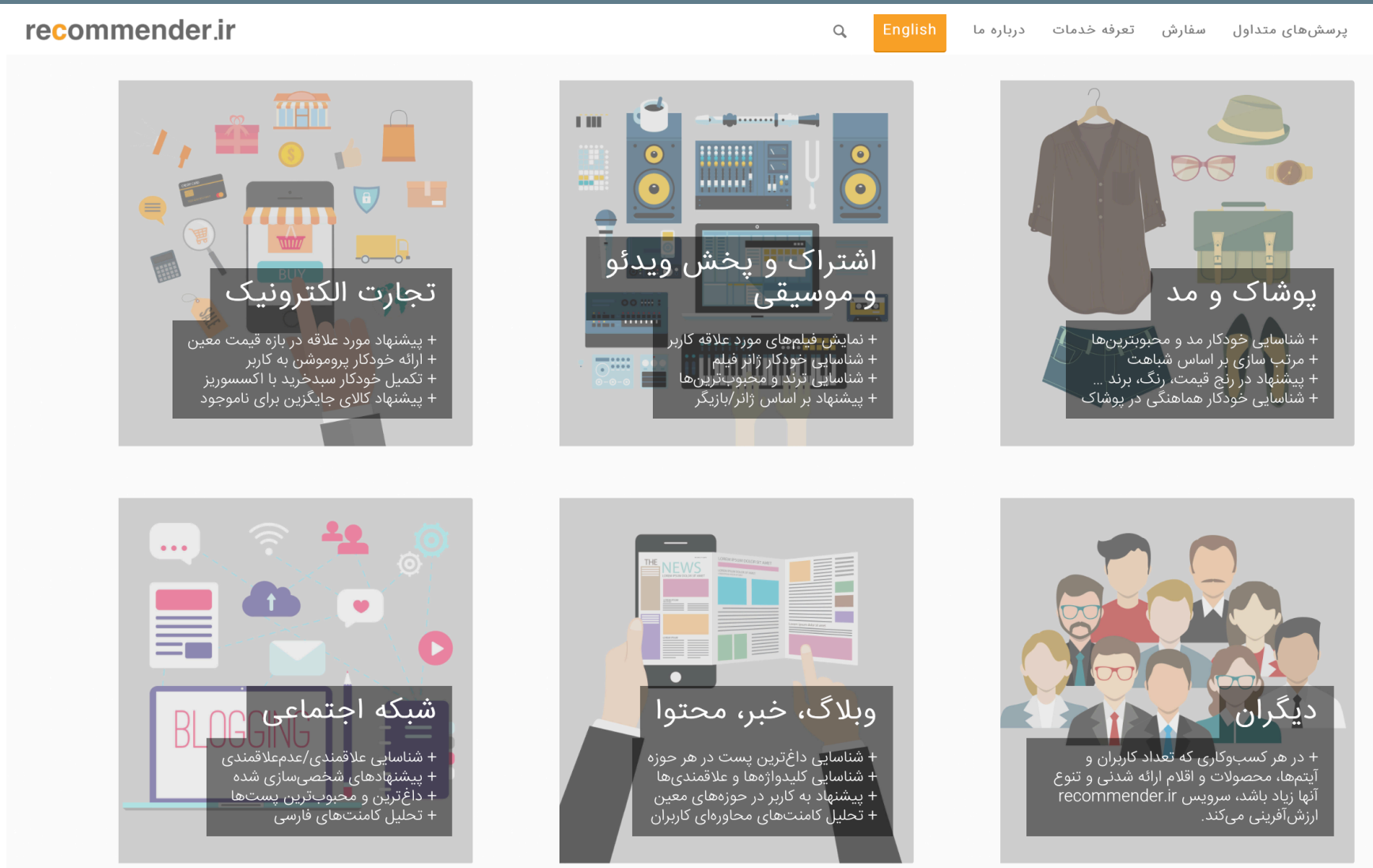


P. Covington, et al. "Deep Neural Networks for YouTube Recommendations ," 2016.

Examples

How does **recommender.ir** work?

Examples


recommender.ir
پرسش‌های متداول
سفارش
تعرفه خدمات
درباره ما
English
پوشاک و مد
+ شناسایی خودکار مد و محبوبترین‌ها
+ مرتب سازی بر اساس شباهت
+ پیشنهاد در رنج قیمت، رنگ، برند ...
+ شناسایی خودکار هماهنگی در پوشاک
اشتراک و پخش ویدئو و موسیقی
+ نمایش فیلم‌های مورد علاقه کاربر
+ شناسایی خودکار ژانر فیلم
+ شناسایی ترند و محبوب‌ترین‌ها
+ پیشنهاد بر اساس ژانر/بازیگر
تجارت الکترونیک
+ پیشنهاد مورد علاقه در بازه قیمت معین
+ ارائه خودکار پروموشن به کاربر
+ تکمیل خودکار سبدخرید با اکسسوریز
+ پیشنهاد کالای جایگزین برای ناموجود
دیگران
+ در هر کسب‌وکاری که تعداد کاربران و آیتم‌ها، محصولات و اقلام ارائه شدنی و تنوع آنها زیاد باشد، سرویس recommender.ir ارزش‌آفرینی می‌کند.
وبلاگ، خبر، محتوا
+ شناسایی داغ‌ترین پست در هر حوزه
+ شناسایی کلیدواژه‌ها و علاقمندی‌ها
+ پیشنهاد به کاربر در حوزه‌های معین
+ تحلیل کامنت‌های محاوره‌ای کاربران
شبکه اجتماعی
+ شناسایی علاقمندی/عدم‌علاقمندی
+ پیشنهادهای شخصی‌سازی شده
+ داغ‌ترین و محبوب‌ترین پست‌ها
+ تحلیل کامنت‌های فارسی

Exercise

One data mining case example from Kaggle in the following format (Post in the telegram group until the next session).

Case	Feature #	Train Sample #	Test Sample #

Kaggle is the number one stop for data science enthusiasts.

Kaggle

kaggle Competitions Datasets Kernels Discussion Jobs ...

Featured Prediction Competition

Porto Seguro's Safe Driver Prediction

Predict if a driver will file an insurance claim next year.

$25,000
Prize Money

Porto Seguro · 5,169 teams · 2 months ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Overview

Description
Evaluation
Prizes
Timeline

Nothing ruins the thrill of buying a brand new car more quickly than seeing your new insurance bill. The sting's even more painful when you know you're a good driver. It doesn't seem fair that you have to pay so much if you've been cautious on the road for years.

Porto Seguro, one of Brazil's largest auto and homeowner insurance companies, completely agrees. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones.

In this competition, you're challenged to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. While Porto Seguro has used machine learning for the past 20 years, they're looking to Kaggle's machine learning community to explore new, more powerful methods. A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

Leaderboard >

1 Michael Jahrer

Kernels >

Porto Seguro Data Exploration
1 vote · 10 hours to go

429 discussion topics >

1st place with representation learni...
213 replies · 16 hours to go

Kaggle

Case	To build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.
Features#	59
Train Sample #	595212
Test Sample #	892816

```
In [1]: import pandas as pd

        train = pd.read_csv('../data/train.csv')
        test = pd.read_csv('../data/test.csv')
```

```
In [2]: train.shape
Out[2]: (595212, 59)
```

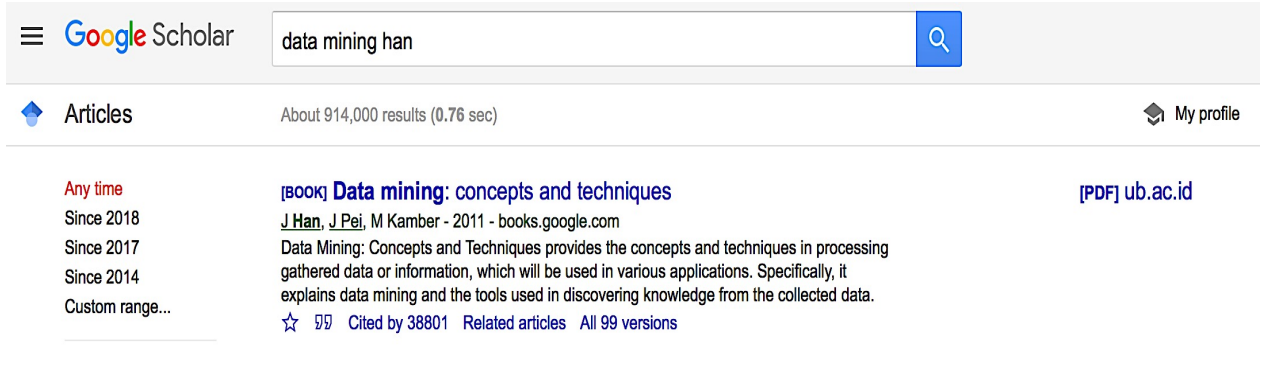
```
In [3]: test.shape
Out[3]: (892816, 58)
```

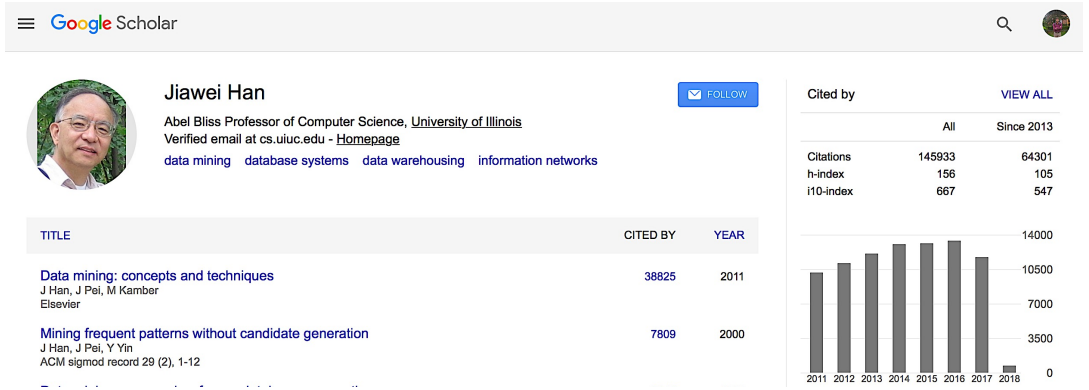
```
In [4]: train.head(5)
```

Out[4]:

	id	target	ps_ind_01	ps_ind_02_cat	ps_ind_03	ps_ind_04_cat	ps_ind_05_cat	ps_ind_06_bin	ps_ind_07_bin	ps
0	7	0	2	2	5	1	0	0	1	0
1	9	0	1	1	7	0	0	0	0	1
2	13	0	5	4	9	1	0	0	0	1
3	16	0	0	1	2	0	0	1	0	0
4	17	0	0	2	0	1	0	1	0	0

Popular websites



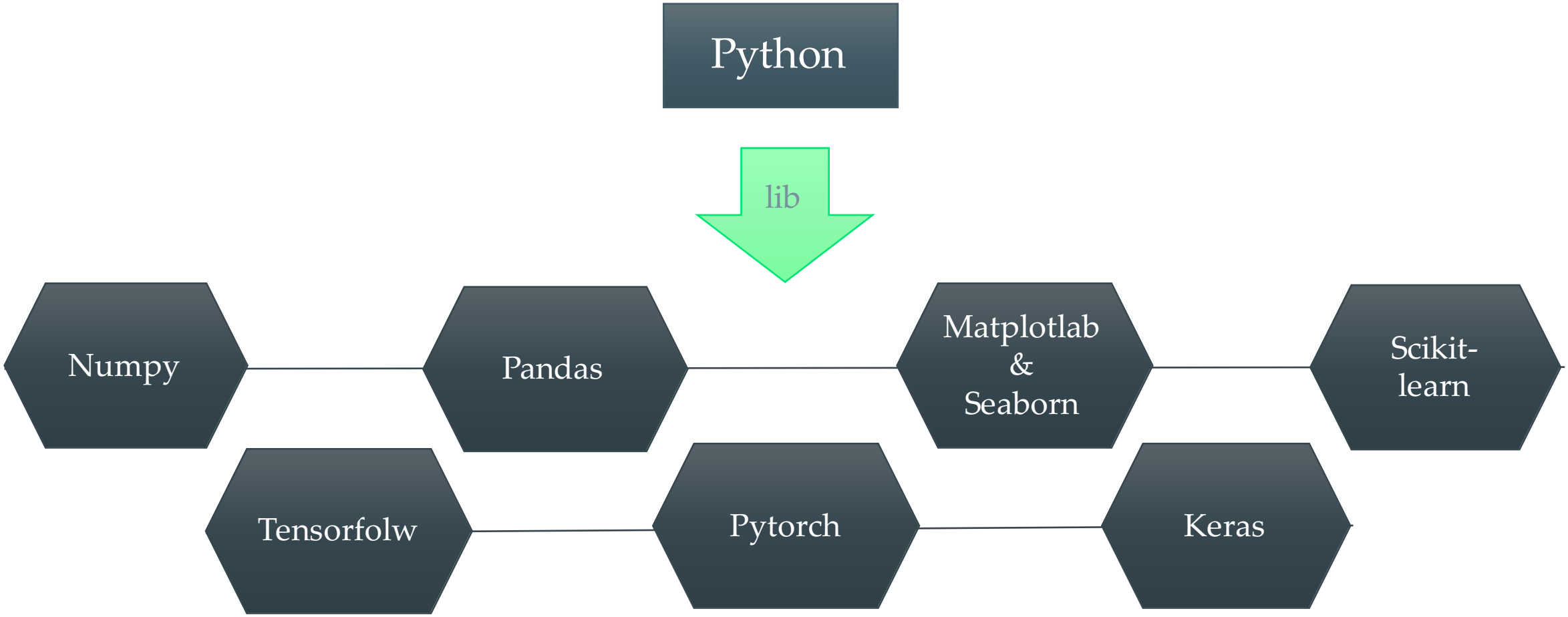


stackoverflow

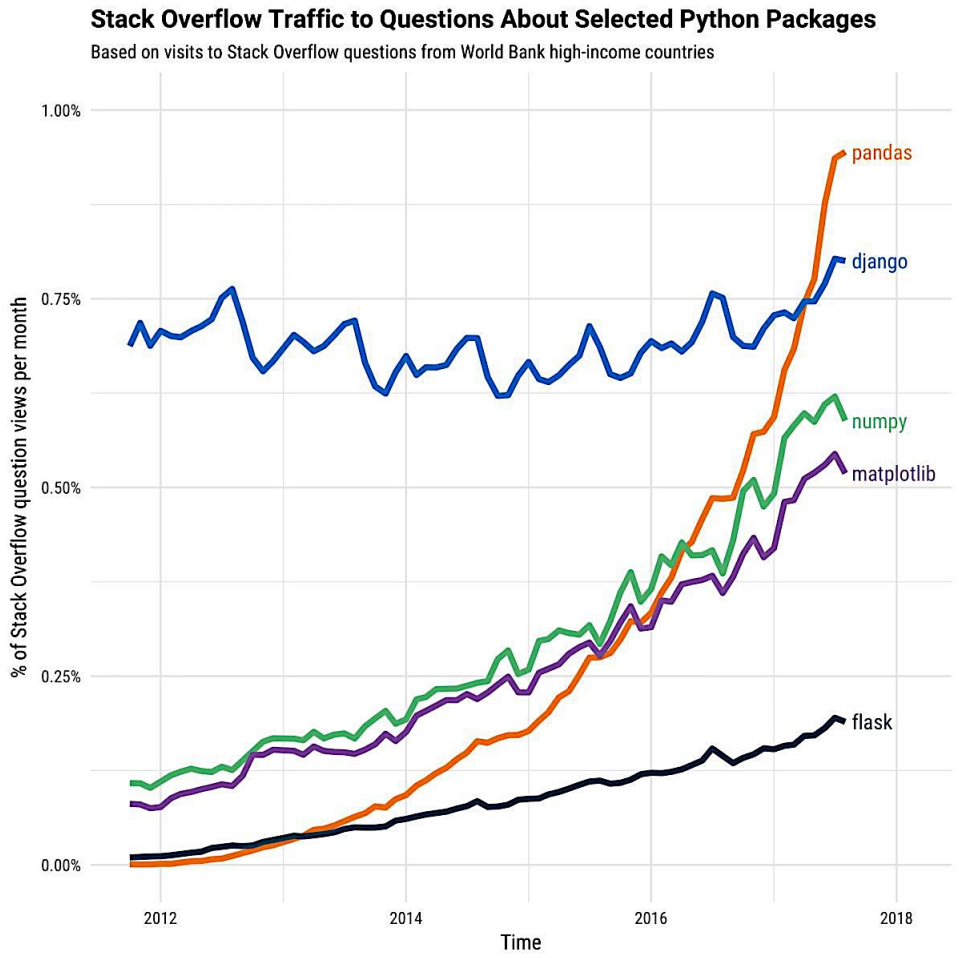
PyData
39,370 subscribers

Tools and Programming language

Tools and Programming language

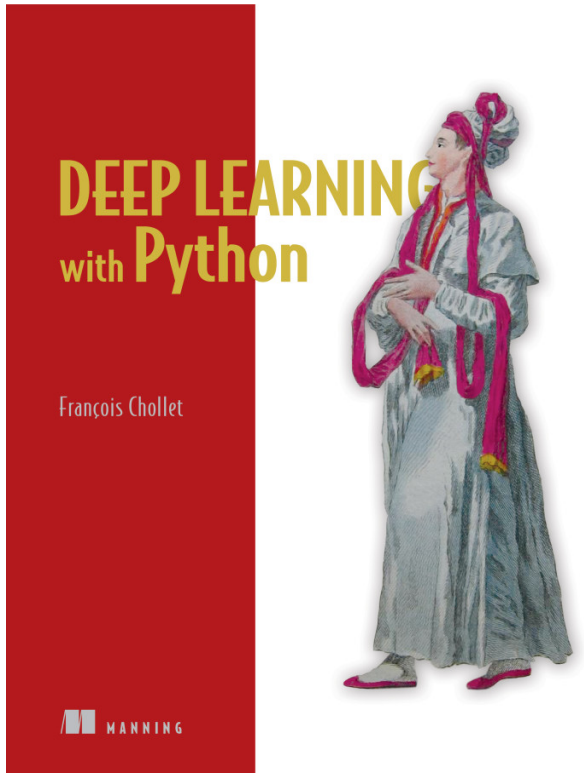


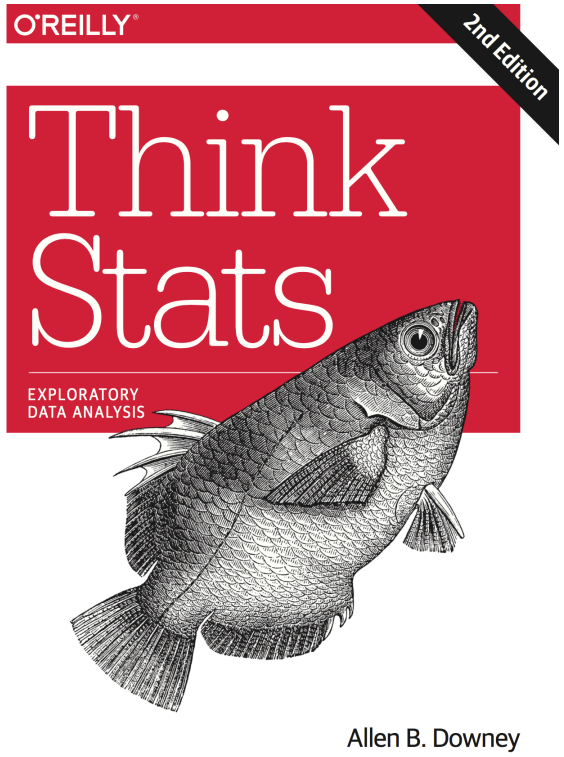
Tools and Programming language

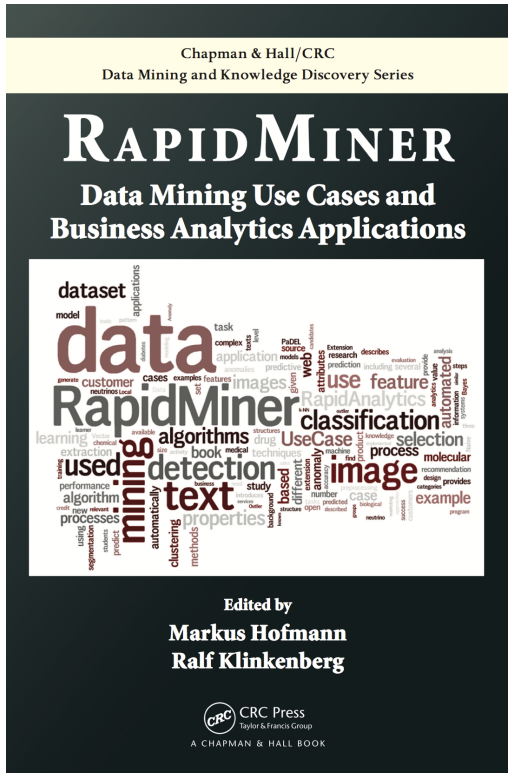


https://stackoverflow.blog/2017/09/14/python-growing-quickly/?cb=1

Resources


DEEP LEARNING
with Python
François Chollet
MANNING


O'REILLY
2nd Edition
Think Stats
EXPLORATORY
DATA ANALYSIS
Allen B. Downey


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
RAPIDMINER
Data Mining Use Cases and
Business Analytics Applications
Edited by
Markus Hofmann
Ralf Klinkenberg
CRC Press
A CHAPMAN & HALL BOOK

Fallacies of data mining

Fallacy 1

There are data mining tools that we can turn loose on our data repositories, and find answers to our problems.

- ***Reality***: data mining is a process.

Fallacy 2

The data mining process is autonomous, requiring little or no human oversight.

- ***Reality***: Data mining is not magic.
- Blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data.
- The wrong analysis is worse than no analysis.

Fallacies of data mining

Fallacy 3 Data mining software packages are intuitive and easy to use.

- ***Reality***: Data analysts must combine subject matter knowledge with an analytical mind, and a familiarity with the overall business or research model.

Fallacy 4 Data mining pays for itself quite quickly.

- ***Reality***: The return rates vary, depending on personnel costs, warehouse costs, etc.

Fallacies of data mining

Fallacy 5 Data mining will identify the causes of our business or research problems.

- ***Reality***: It is up to the humans to identify the causes.

Fallacy 6 Data mining will automatically clean up our messy database.

- ***Reality***: Not automatically.

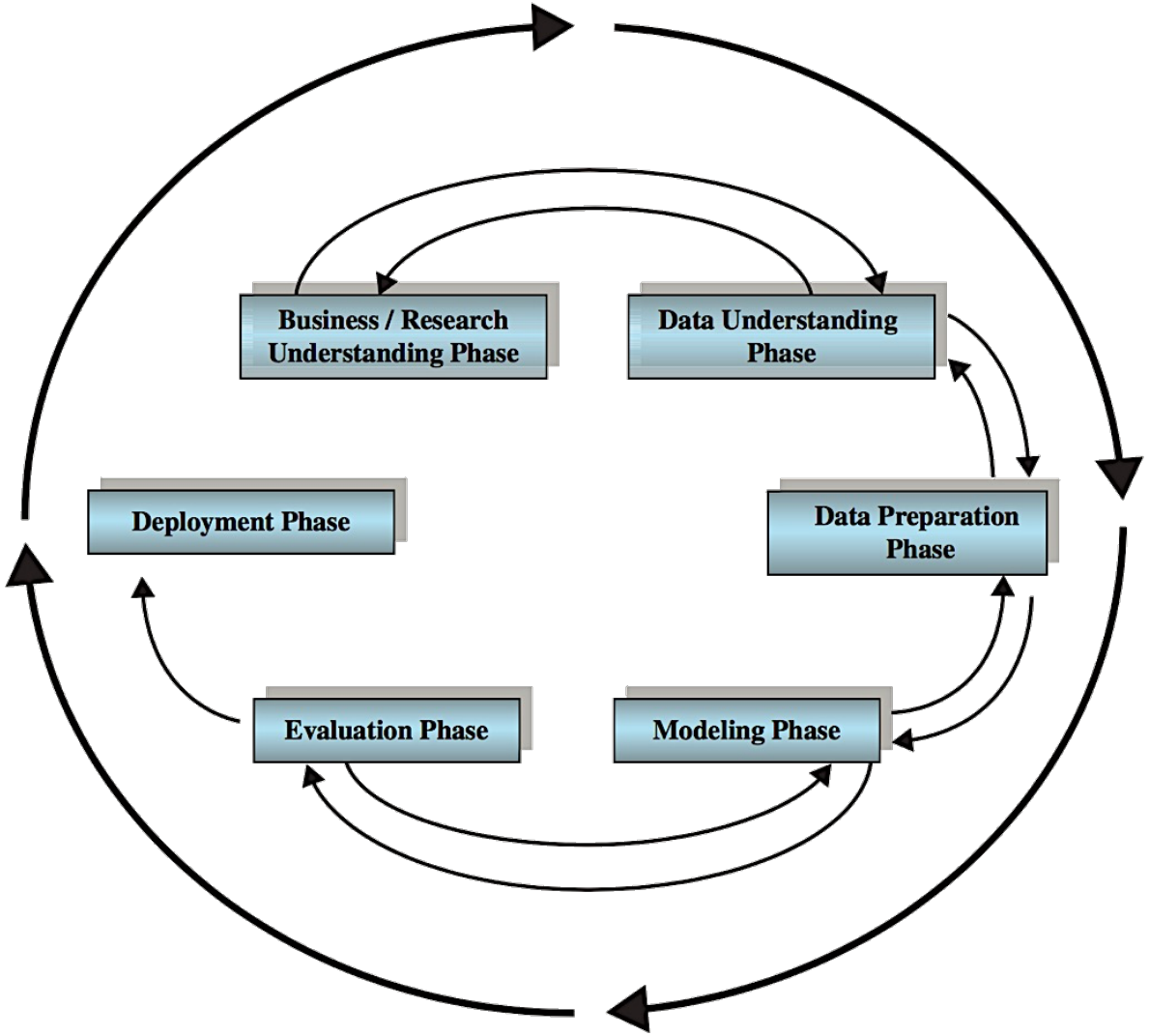
Standard process for data mining

- A cross-industry standard is clearly required, that is industry-neutral, tool-neutral, and application-neutral.

- Wikipedia: Polls conducted at one and the same website (KDNuggets) in 2002, 2004, 2007 and 2014 show that CRISP-DM was the leading methodology used by industry data miners who decided to respond to the survey.

- **CRISP-DM: Cross-Industry Standard Process for Data Mining.**

CRISP-DM


Business / Research Understanding Phase
Data Understanding Phase
Data Preparation Phase
Modeling Phase
Evaluation Phase
Deployment Phase

CRISP-DM

1. **Business/Research Understanding Phase**
   - Clearly enunciate the project objectives and requirements.
   - Translate these goals into the formulation of a data mining problem.
   - Prepare a preliminary strategy for achieving these objectives.
2. **Data Understanding Phase**
   - Collect the data.
   - Use exploratory data analysis to familiarize yourself with the data, and discover initial insights.
   - Evaluate the quality of the data.
   - Select interesting subsets that may contain actionable patterns.
3. **Data Preparation Phase**
   - This labor-intensive phase covers all aspects of preparing the final data set, from the initial, raw, dirty data.
   - Select the cases and variables appropriate for your analysis.
   - Perform transformations on certain variables, if needed.
   - Clean the raw data so that it is ready for the modeling tools.
4. **Modelling Phase**
   - Select and apply appropriate modeling techniques.
   - Calibrate model settings to optimize results.
   - May require looping back to data preparation phase, in order to bring the form of the data into line with data mining technique.
5. **Evaluation Phase**
   - These models must be evaluated for quality and effectiveness.
   - Determine whether the model in fact achieves the objectives set for it in Phase 1.
   - Finally, come to a decision regarding the use of the data mining results.
6. **Deployment Phase**
   - Example of a simple deployment: Generate a report.
   - More complex: Implement a parallel data mining process in another department.
   - For businesses, the customer often carries out the deployment based on your model.

Note

Build your first system quickly, then iterate
–Andrew Ng

Premature optimization is the root of all evil.
– Donald Knuth

Major tasks

- Description
- Prediction
- Regression
- Clustering
- Classification
- Association

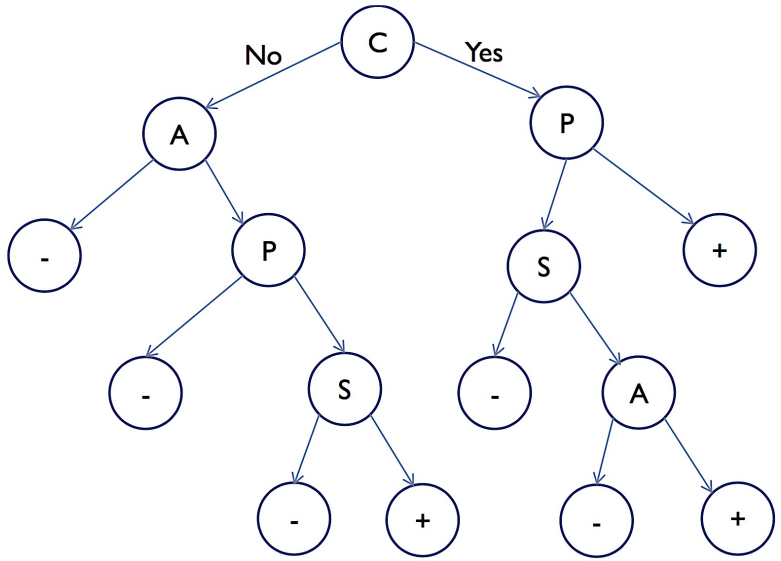
- The obtained model from data is used as
  - Discovered knowledge from data.
  - Tool for deciding about samples not available in training data.

Data mining task 1: Description

- To describe patterns and trends lying within the data.

- Data mining models should be as **transparent** as possible.
  - For example, **decision trees** provide an intuitive and human-friendly explanation.
  - On the other hand, **neural networks** are comparatively opaque to nonspecialists.

- Attributes:
  - A: age>40
  - C: chest pain
  - S: smoking
  - P: physical test

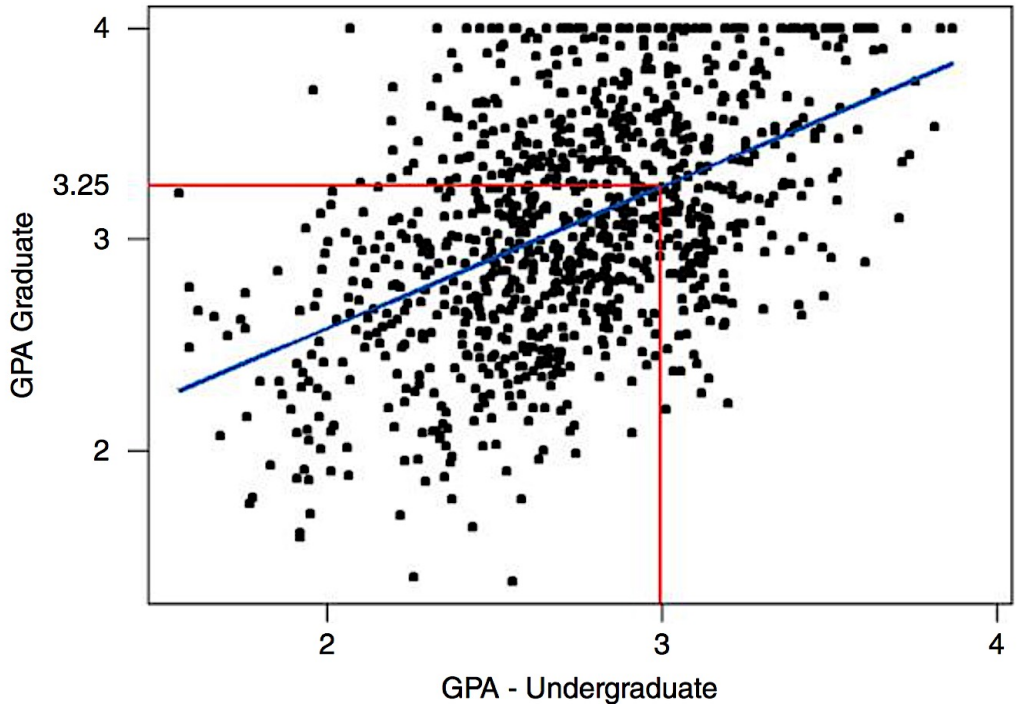


- Label:
  - Heart disease (+), No heart disease (-)

Ref: Dr. Soleymani ML slides.

Data mining task 2: Regression

- Approximate the value of a **numeric** target variable using a set of numeric and/or categorical predictor variables.
- Models are built using a **training set**.
- Example: Estimating the grade point average (GPA) of a graduate student, based on that student's undergraduate GPA.



Graphs and plots are helpful for understanding two and three dimensional relationships in data, but are not enough for multi-dimensional settings.

Data mining task 3: Classification

- Similar to regression, except that the target variable is categorical rather than numeric. Such as
  - Determining whether a particular credit card transaction is fraudulent.
  - Diagnosing whether a particular disease is present.

- The most important current success of data mining and machine learning.

Data mining task 4: Prediction

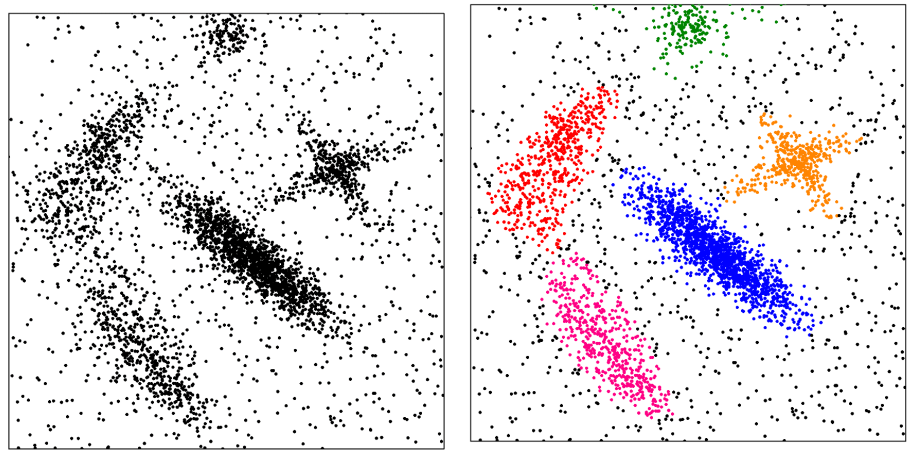
- Prediction is similar to classification and estimation, except that for prediction, the results lie in the future.

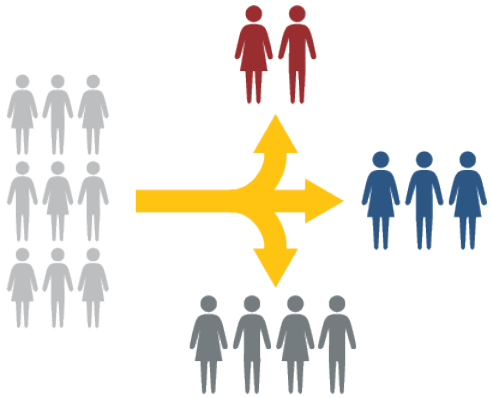
- Examples:
  - Predicting the price of a stock 3 months into the future.
  - Predicting the percentage increase in traffic deaths next year if the speed limit is increased.

- Any of the methods used for classification and estimation may also be used, under appropriate circumstances, for prediction.

Data mining task 5: Clustering

- The grouping of records, observations, or cases into classes of similar objects.
- A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters.
- The clustering task does not try to classify, estimate, or predict the value of a target variable.
- Clustering is often performed as a preliminary step in a data mining process.





Data mining task 6: Association

- The job of finding which attributes "go together."
- Most prevalent in the business world, where it is known as **affinity analysis** or **market basket analysis**.
- The rules are in the form: "*If antecedent then consequent*".
- Examples:
  - Examining the proportion of children whose parents read to them who are themselves good readers
  - Finding out which items in a supermarket are purchased together, and which items are never purchased together.

References

D. T. Larose, C. D. Larose, “Discovering knowledge in data,” Wiley, 2014. Chapter 1.

J. Han, M. Kamber, J. Pei, “Data mining, concepts and techniques,” 3rd edition, Morgan Kaufmann, 2011.

The End