

Foundations of Data Science & Analytics: Course Overview

Ezgi Siir Kibris

[Introduction to Data Mining, 2nd Edition](#)

by

Tan, Steinbach, Karpatne, Kumar

Instructor

- Ezgi Siir Kibris: eskics@rit.edu
- Office: GOL 2669
- Office hours:

Schedule an appointment via Calendly:

<https://calendly.com/eskics-g>

If these times do not work for you, please let me know.

Monday/Wednesday: 12:00 pm-1:00 pm

Friday: 12:00 am-1:00 pm and 3:00-4:00pm

Zoom: <https://rit.zoom.us/j/8189920424>

Github

- Syllabus
- Assignments
- Slides

<https://github.com/ezgisiir/fds>

Assignment 0

- Create a Github account
- Write it to Google sheet

Data Mining is Everywhere



Cyber Security



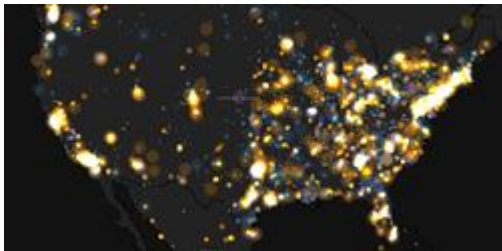
Traffic Patterns



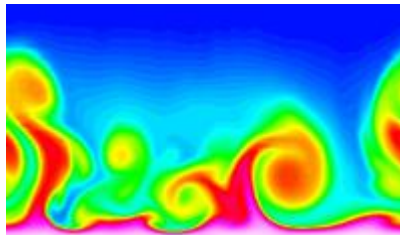
Sensor Networks



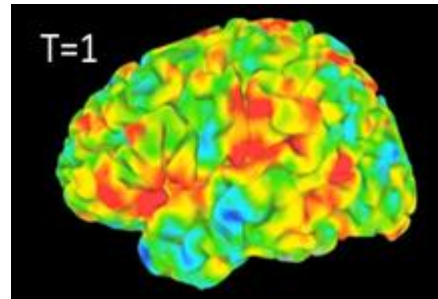
E-Commerce



Social Networking: Twitter



Computational Simulations



Brain Activity

Behind the Trend (Why?)

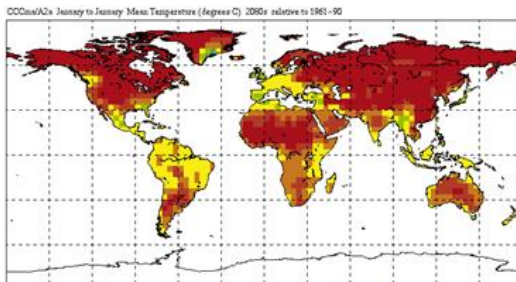
- Lots of data is being collected and warehoused
 - Yahoo has Peta Bytes of web data
 - Facebook has billions of active users
 - Amazon handles millions of visits/day
- Computers have become cheaper and more powerful



Great Opportunity



Improving health care and reducing costs



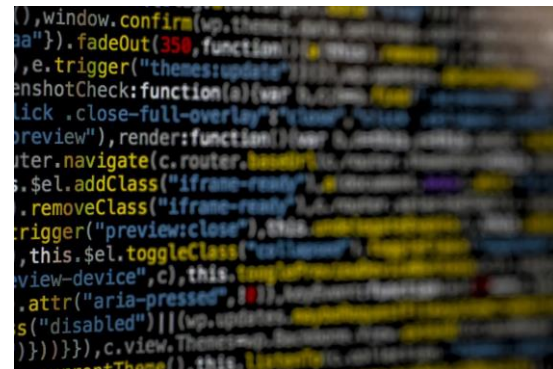
Predicting the impact of climate change



Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production



Optimizing the development of software

What

What is data mining?

What is NOT Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

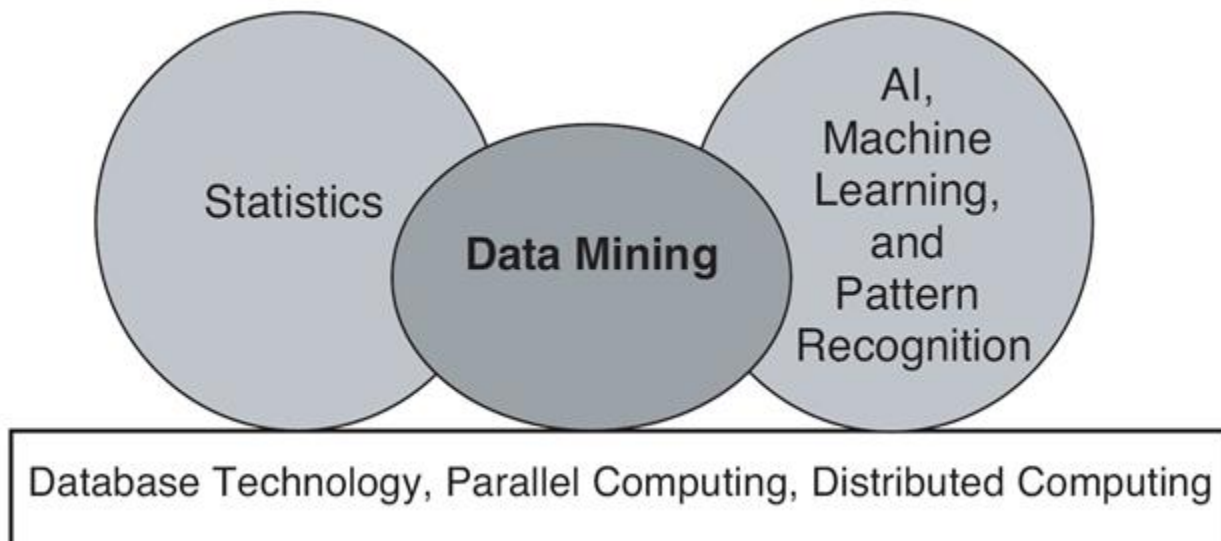
What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Predict whether user want information on Amazon rainforest or Amazon.com when searching for “Amazon”

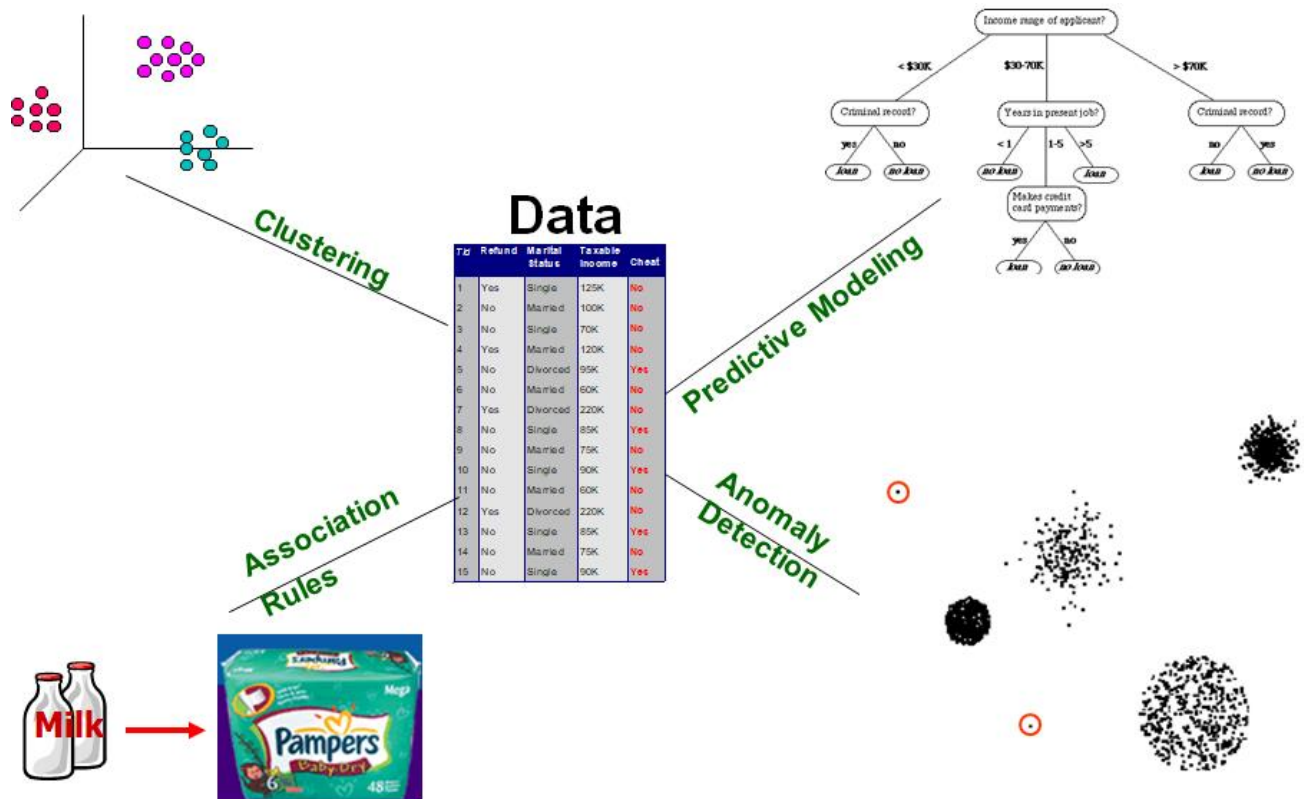
Origin of Data Mining

Especially for data:

- Large-scale
- High dimensional
- Complex
- Distributed



Tasks



01/17/2018

Introduction to Data Mining, 2nd Edition

11

Predictive Modeling (supervised learning)

**Independent Variable
(Input)**



Cat

Train:



Dog

Predict:



?

Predictive Modeling

	Output
Classification:	Classes / Categories
Regression:	Continuous Values

Predictive Modeling: Classification

Image Classification



Cat
Dog
Bird
Pig
Lion
...

Predictive Modeling: Classification

Sentiment Analysis (NLP)

**You have been
working on this
for months.
I need to see your
results, now!**



Positive

Neutral

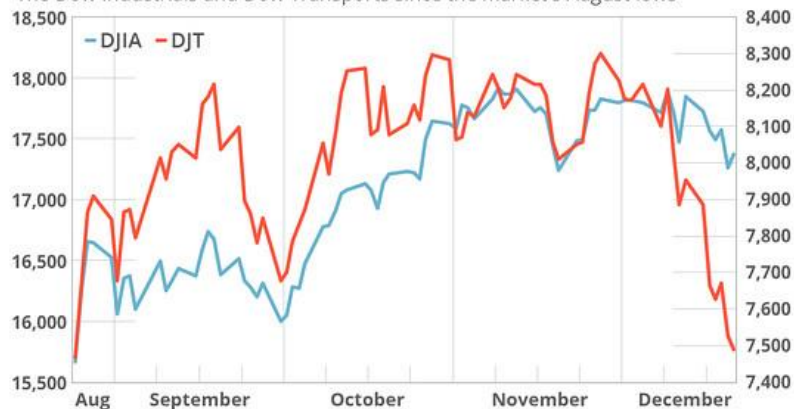
Negative

Predictive Modeling: Regression

Time Series Analysis

The Transports have broken down

The Dow Industrials and Dow Transports since the market's August lows



Source: The Hulbert Financial Digest



**What will
DJIA and
DJT be in
January?**

Predictive Modeling: Regression

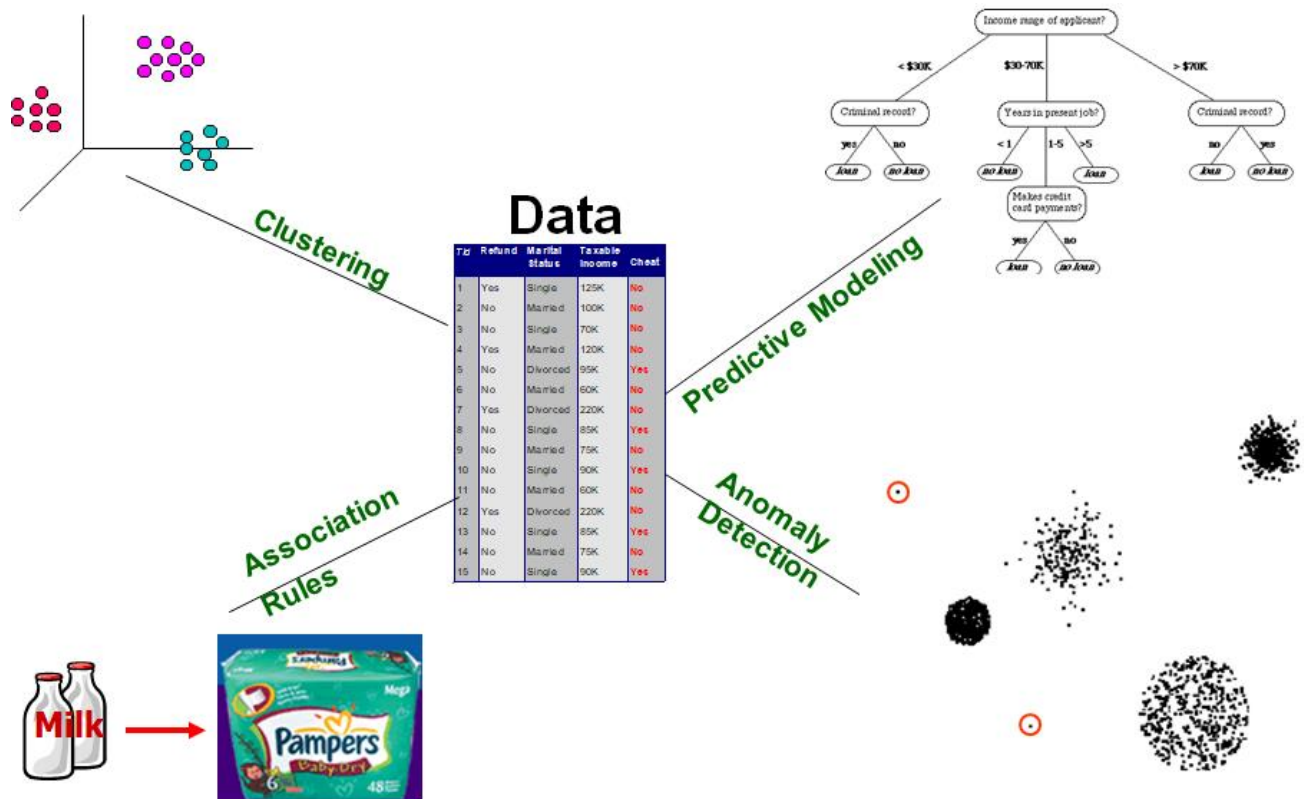
ID	Name	Gender	GRE	Coding
001	Emily Wang	F	320	Yes
002	James Bond	M	320	Yes

Discrimination
Problem?



**Grade in
DSCI-633**

Tasks



01/17/2018

Introduction to Data Mining, 2nd Edition

11

Clustering (unsupervised learning)

Independent Variable

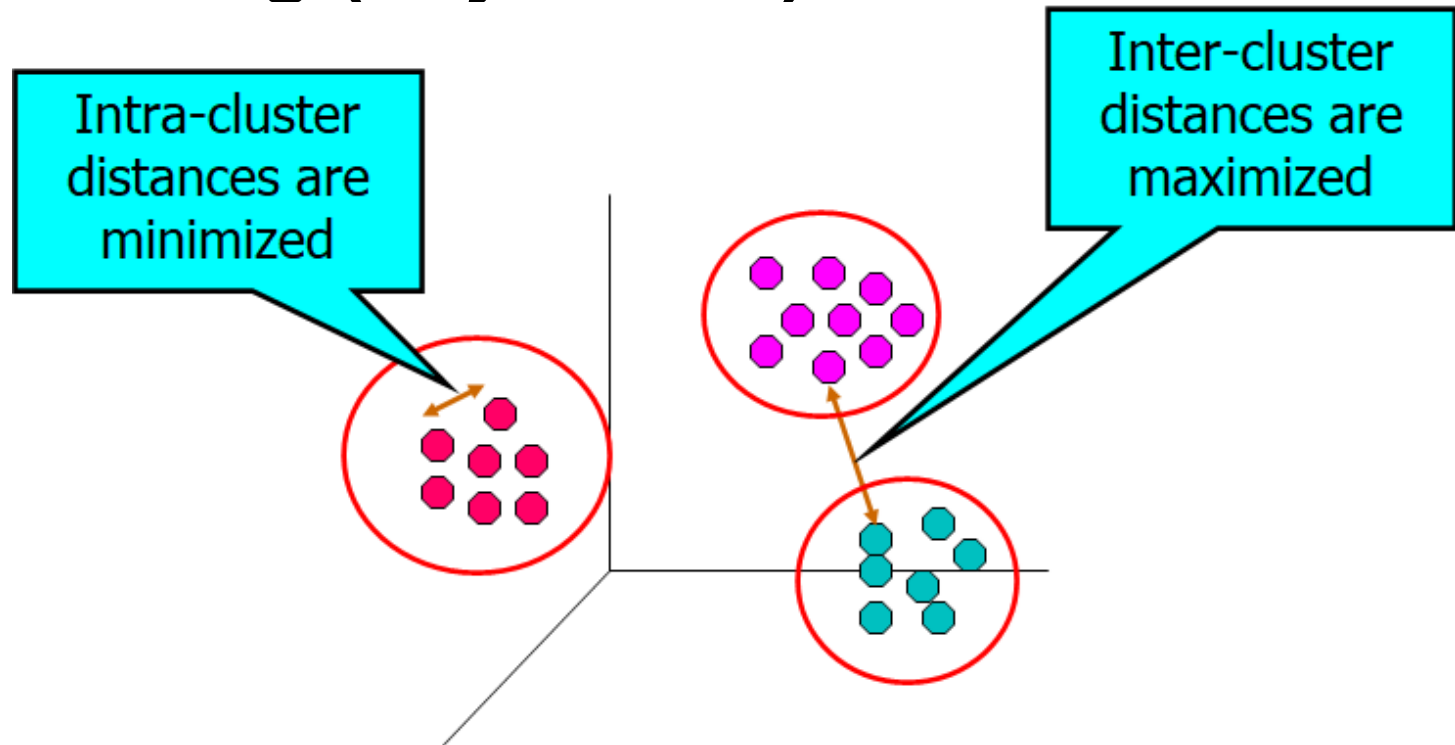


Clustering (unsupervised learning)

Independent Variable



Clustering (objectives)



Clustering (applications)

User ID	Titanic	Die Hard	Avatar
001	Yes	Yes	No
002	Yes	No	Yes



**Groups of
users that
like similar
movies**

Clustering (applications)

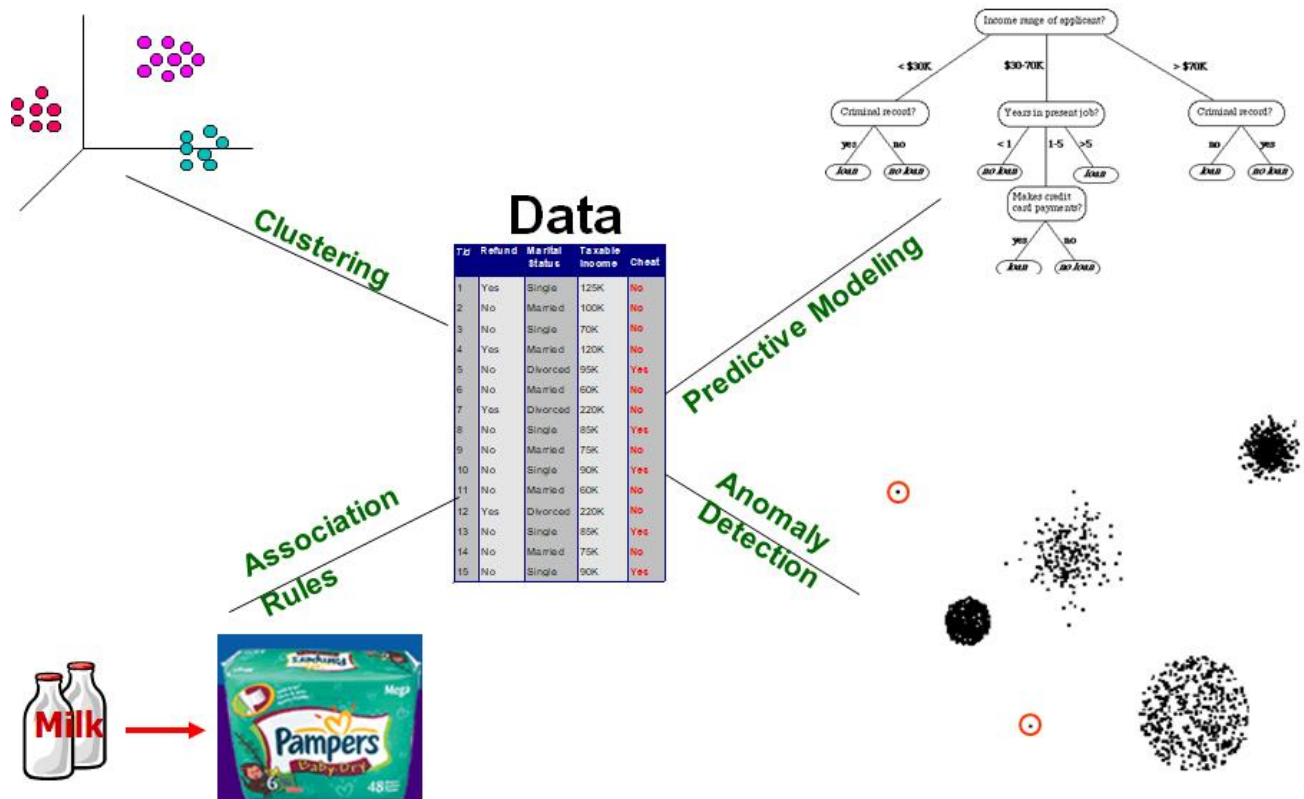
Topic Modeling

Area soccer players
earn all-state
selections.



Sports	Politics	News	Movie
0.4	0.18	0.4	0.02

Tasks



01/17/2018

Introduction to Data Mining, 2nd Edition

11

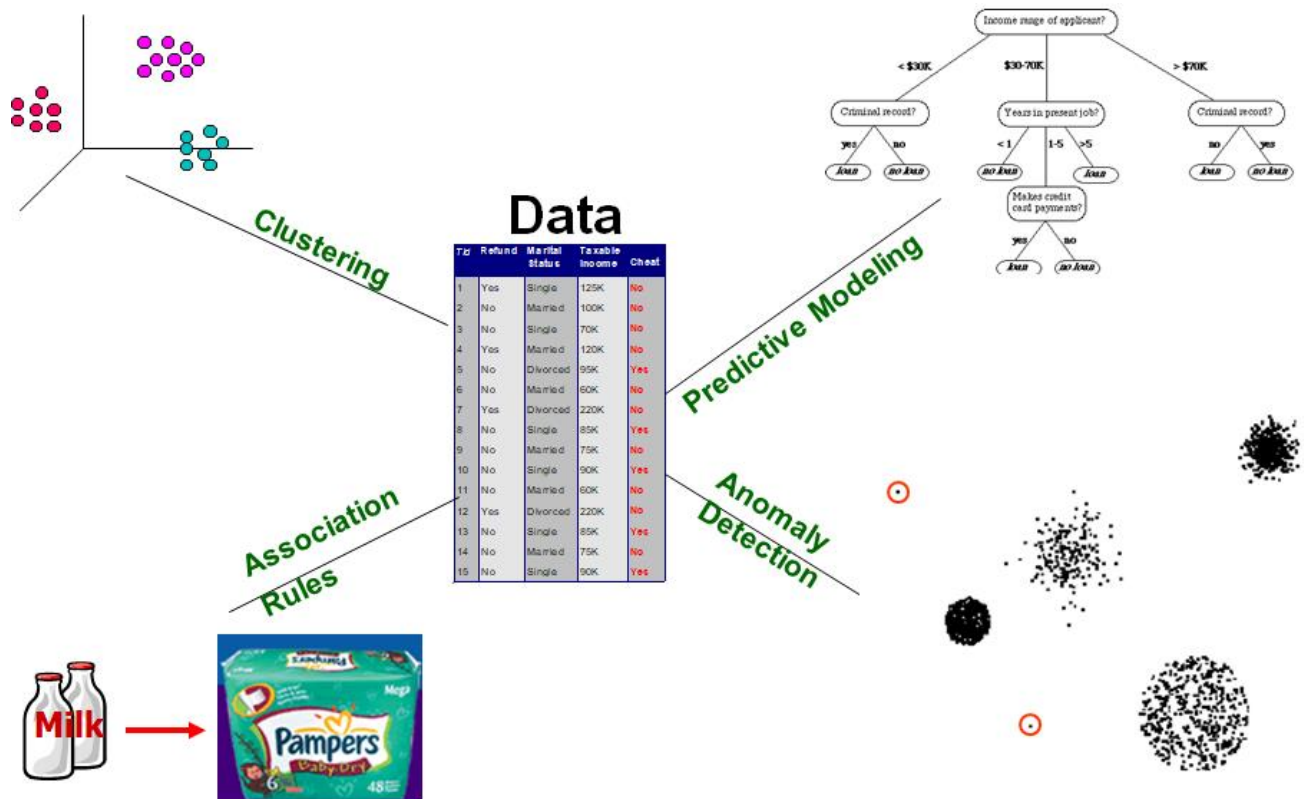
Association Rule Mining

User ID	Titanic	Die Hard	Avatar
001	Yes	Yes	Yes
002	Yes	No	Yes



**People who
watched
Titanic will
also watch
Avatar**

Tasks



01/17/2018

Introduction to Data Mining, 2nd Edition

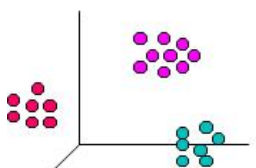
11

Anomaly Detection

Purchase	11/03/2019	11/13/2019	11/23/2019	12/03/2019	12/04/2019
001	Titanic	Avatar	Die Hard	Aliens	Hunter X Hunter

Anomaly Detection Tasks

- Credit card fraud detection
- Network intrusion detection



Clustering

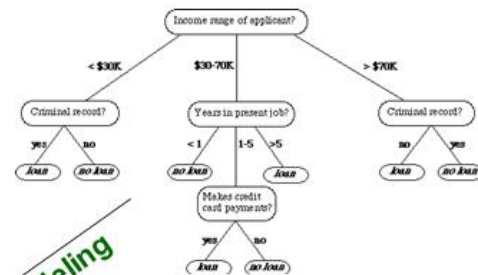
Data

id	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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Predictive Modeling

Anomaly Detection



01/17/2018

Introduction to Data Mining, 2nd Edition

11