School of Computing
UNIVERSITY OF GEORGIA

CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

August 17, 2023

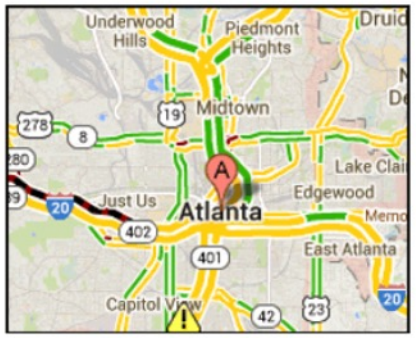
Introduction to Data Mining

Why Data Mining

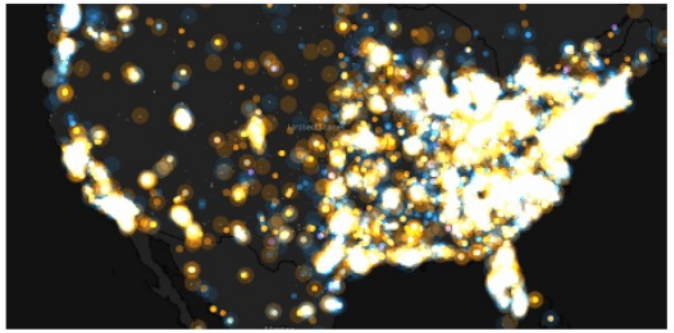
Cyber Security

E-Commerce

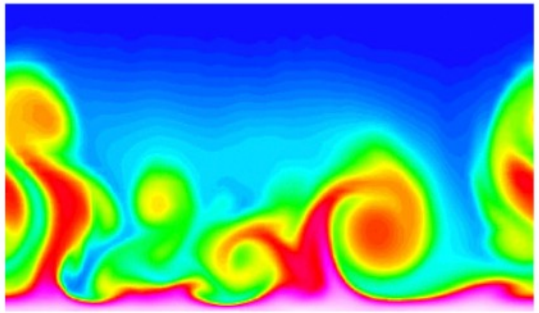
Sensor Networks



Traffic Patterns



Social Networking: Twitter



Computational Simulations

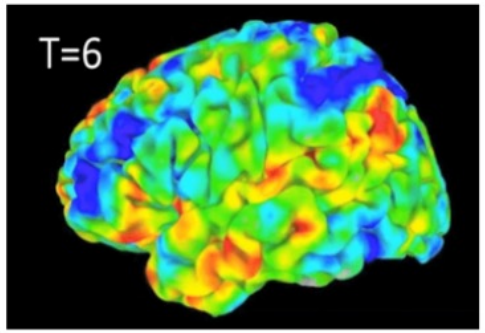
Why Data Mining

- Commercial Viewpoint

  - Lots of data is being collected and warehoused.
  - Computers have become cheaper and more powerful.
  - Competitive pressure is strong.

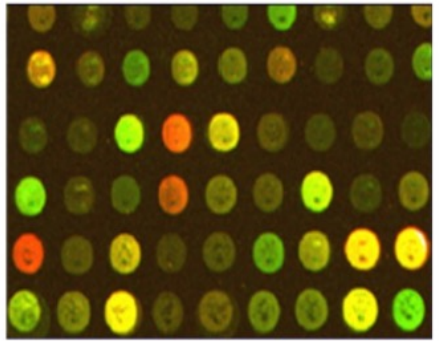
Why Data Mining

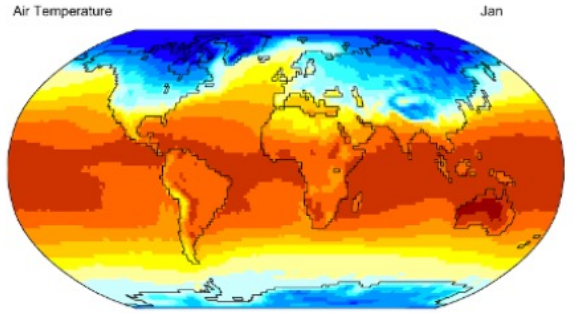
- Scientific Viewpoint



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



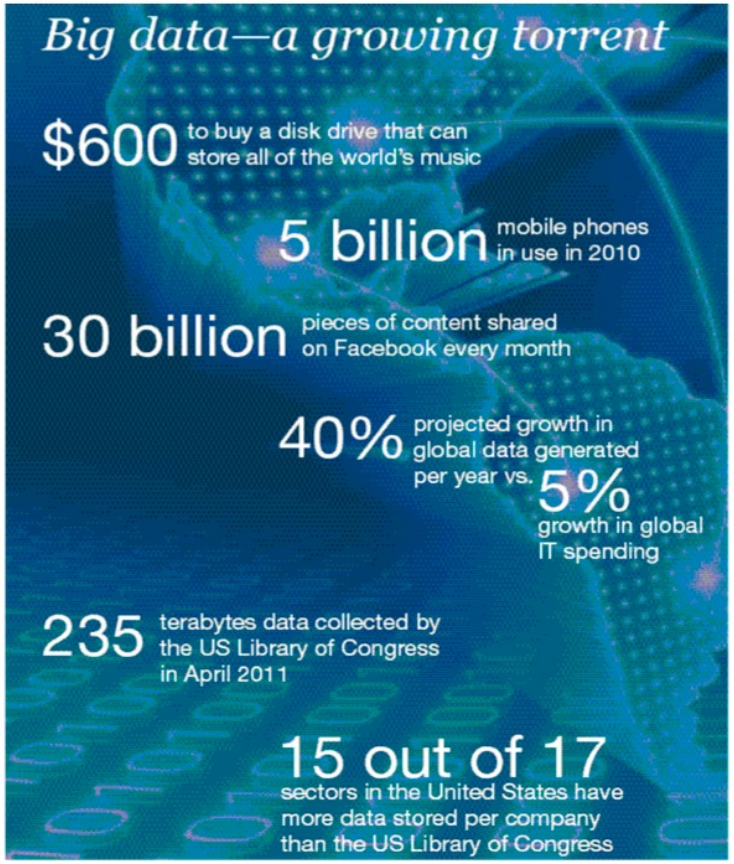
Surface Temperature of Earth

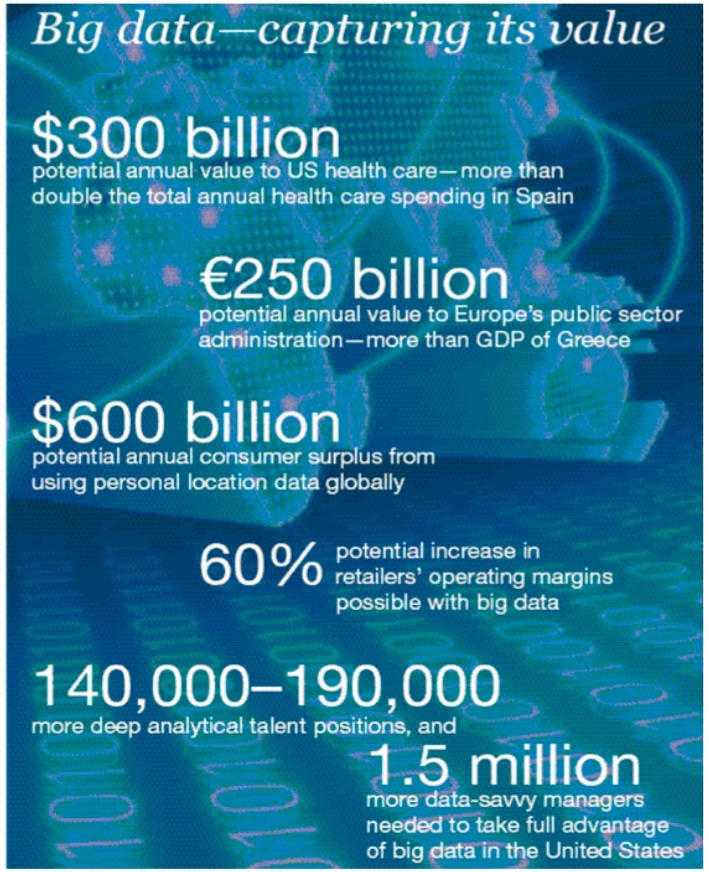
Data collected and stored at enormous scale and speed

- fMRI for patients.
- Telescopes scanning the skies
- High-throughput biological data
- Scientific simulations

Why Data Mining

- Great opportunities to improve productivity.
- Great opportunities to solve society's major problems.

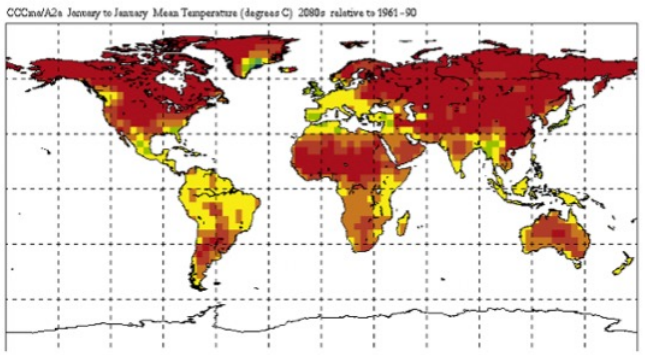




Why Data Mining

- Great opportunities to improve productivity.
- Great opportunities to solve society's major problems.

Improving health care and reducing costs

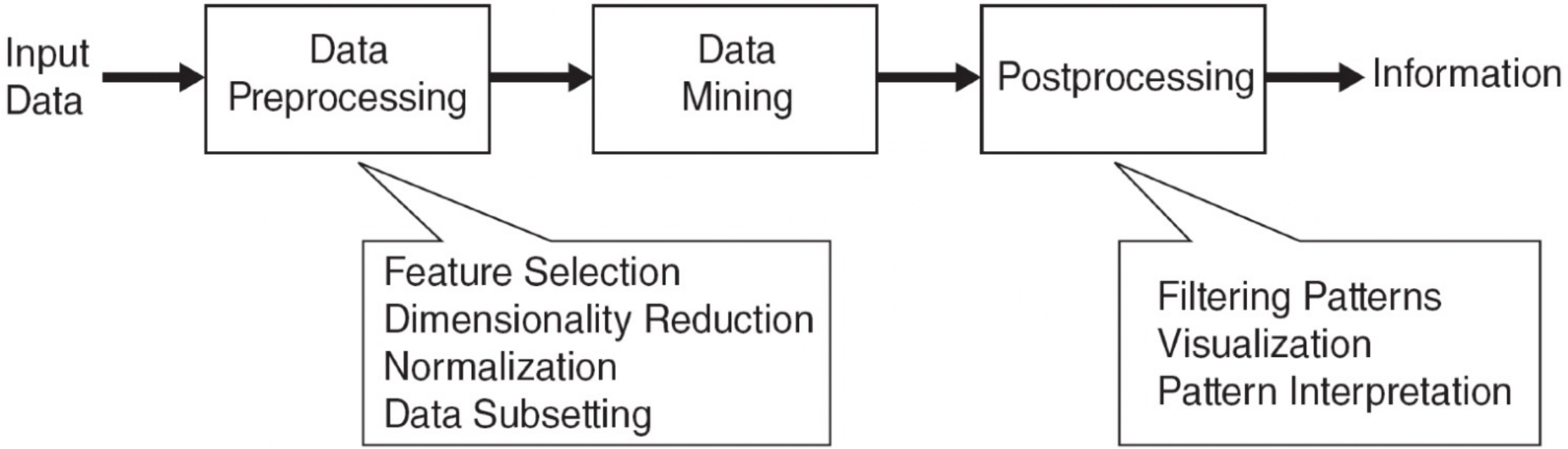


Predicting the impact of climate change

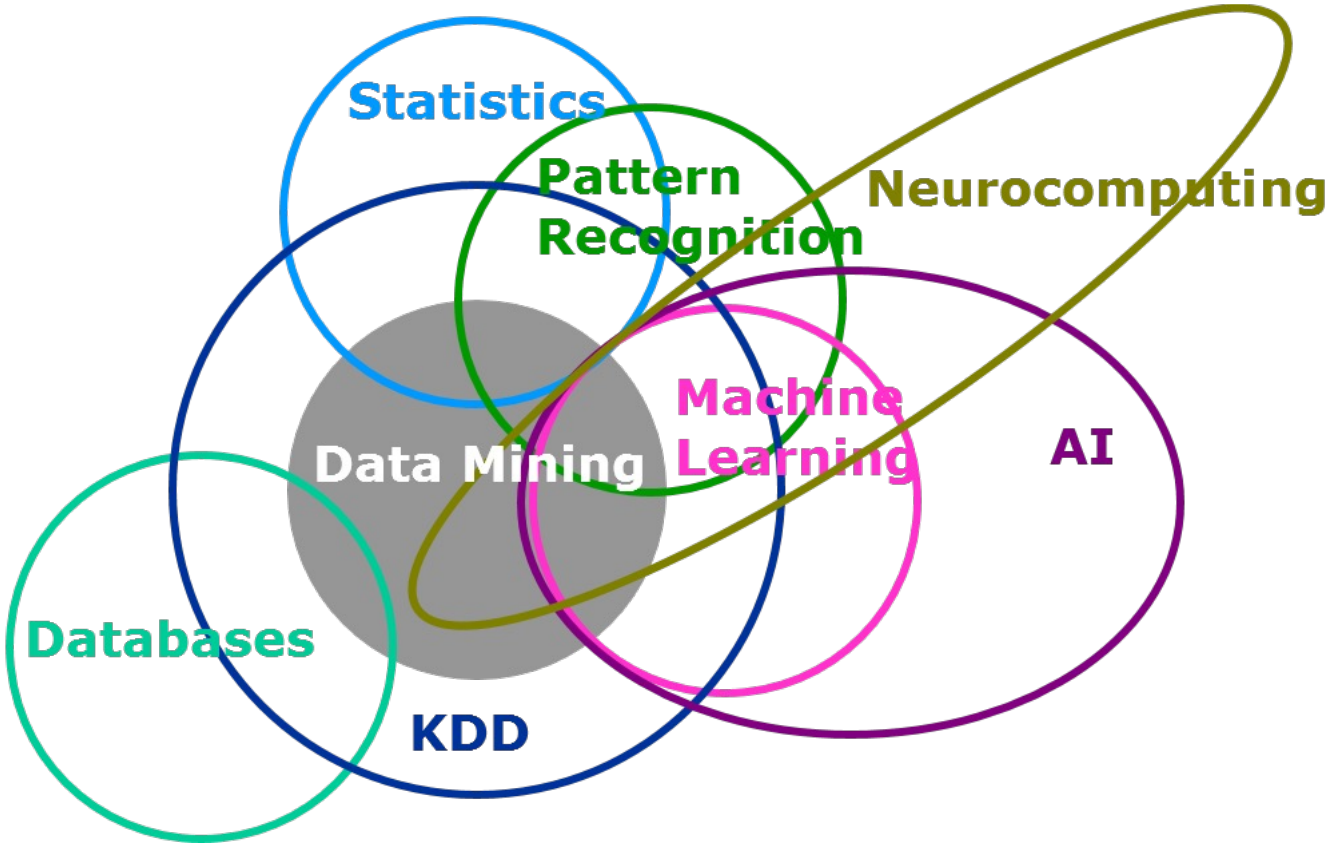
Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

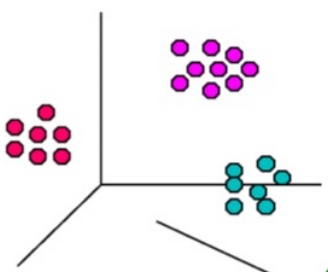
Data Mining Process



Data Mining vs Other Concepts



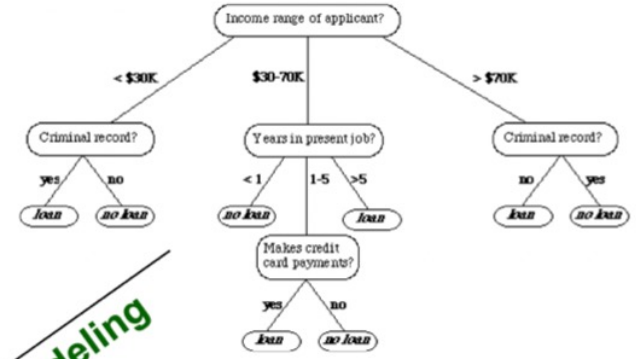
Data Mining



Clustering

Data

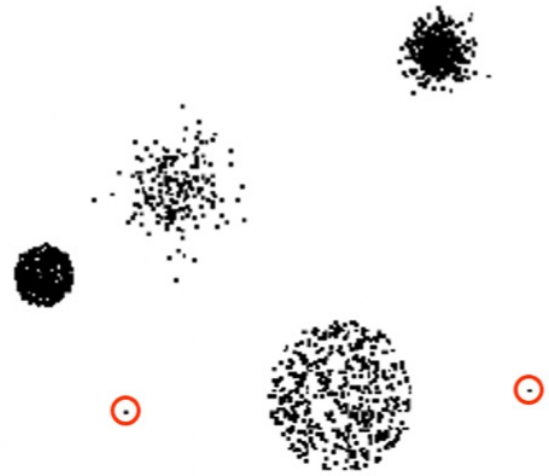
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



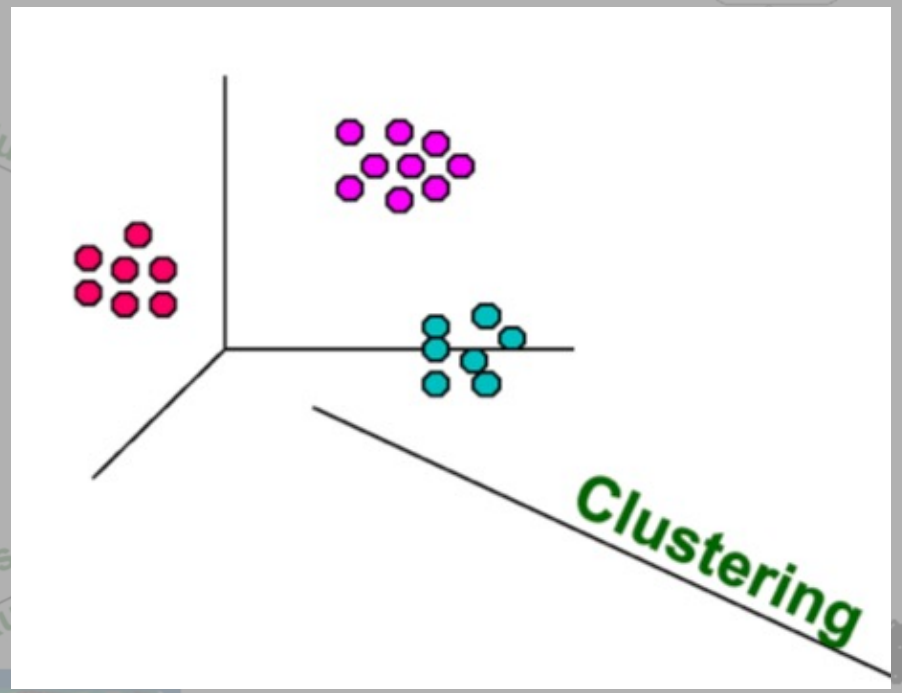
Predictive Modeling

Association Rules

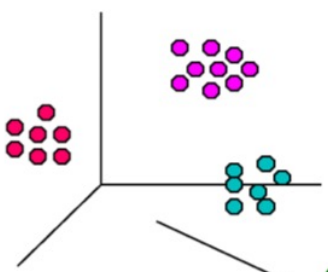
Anomaly Detection



Data Mining



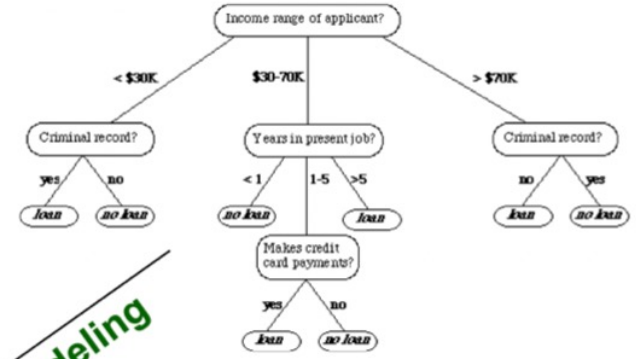
Data Mining



Clustering

Data

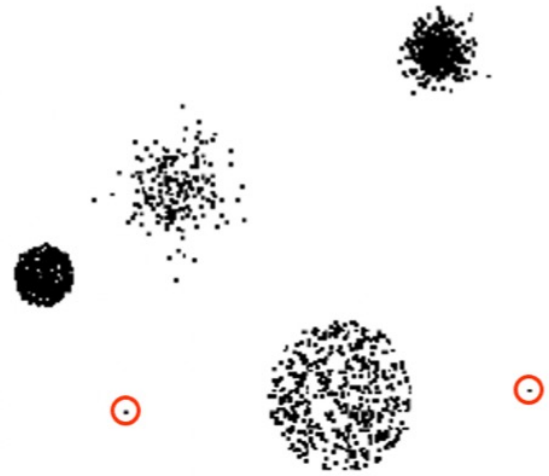
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



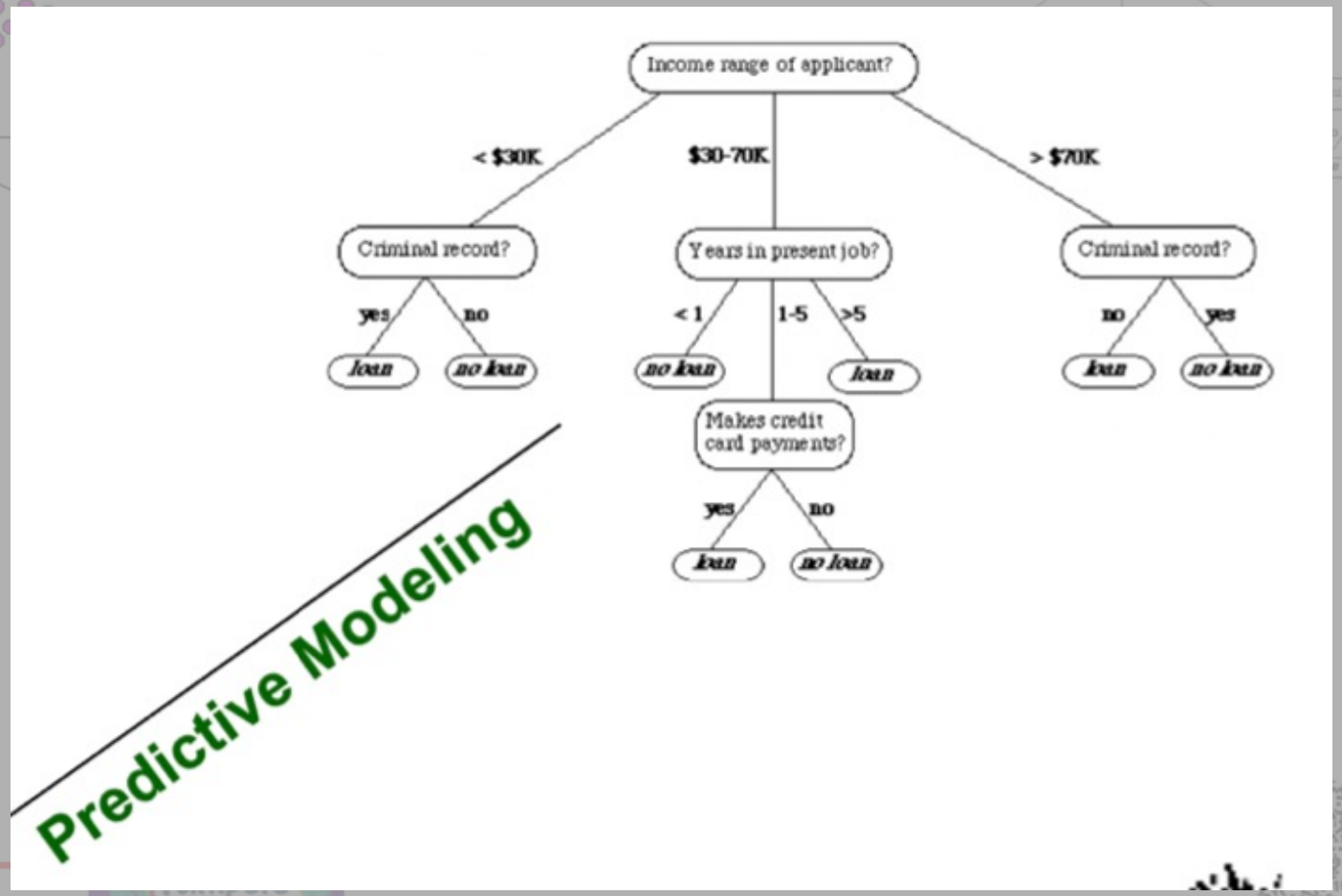
Predictive Modeling

Association Rules

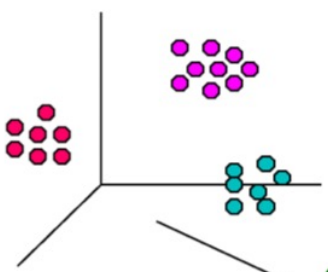
Anomaly Detection



Data Mining


Income range of applicant?
< $30K
$30-70K
> $70K
Criminal record?
Years in present job?
Criminal record?
yes
no
< 1
1-5
>5
no
yes
loan
no loan
no loan
loan
loan
no loan
Makes credit card payments?
yes
no
loan
no loan
Predictive Modeling

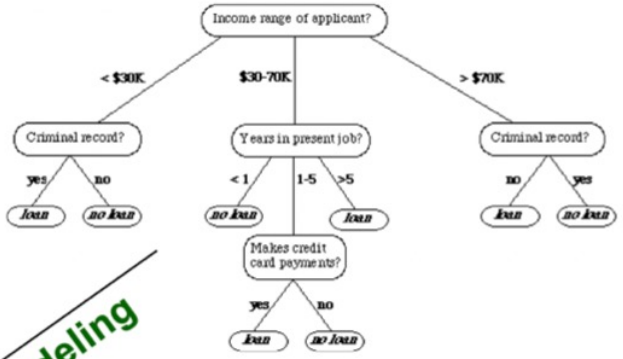
Data Mining



Clustering

Data

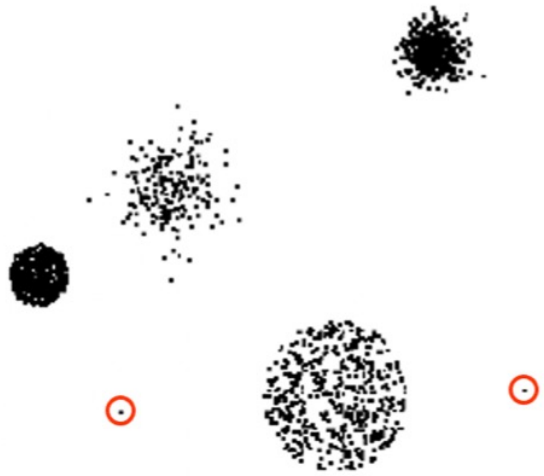
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



Predictive Modeling

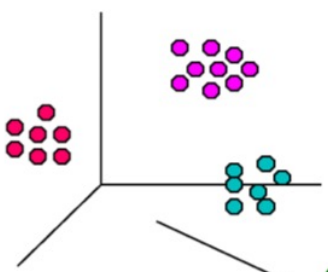
Association Rules

Anomaly Detection



Data Mining

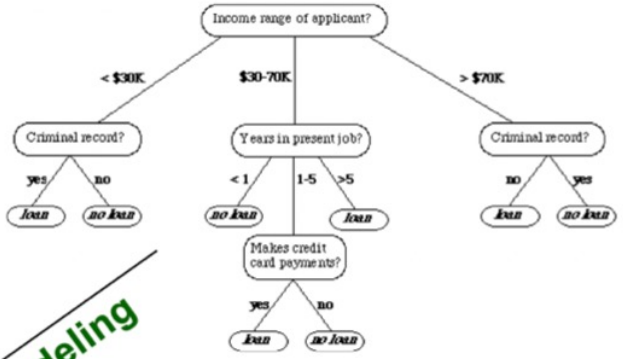
Data Mining



Clustering

Data

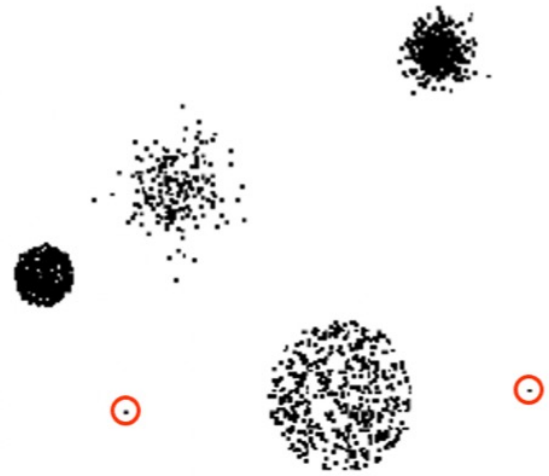
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



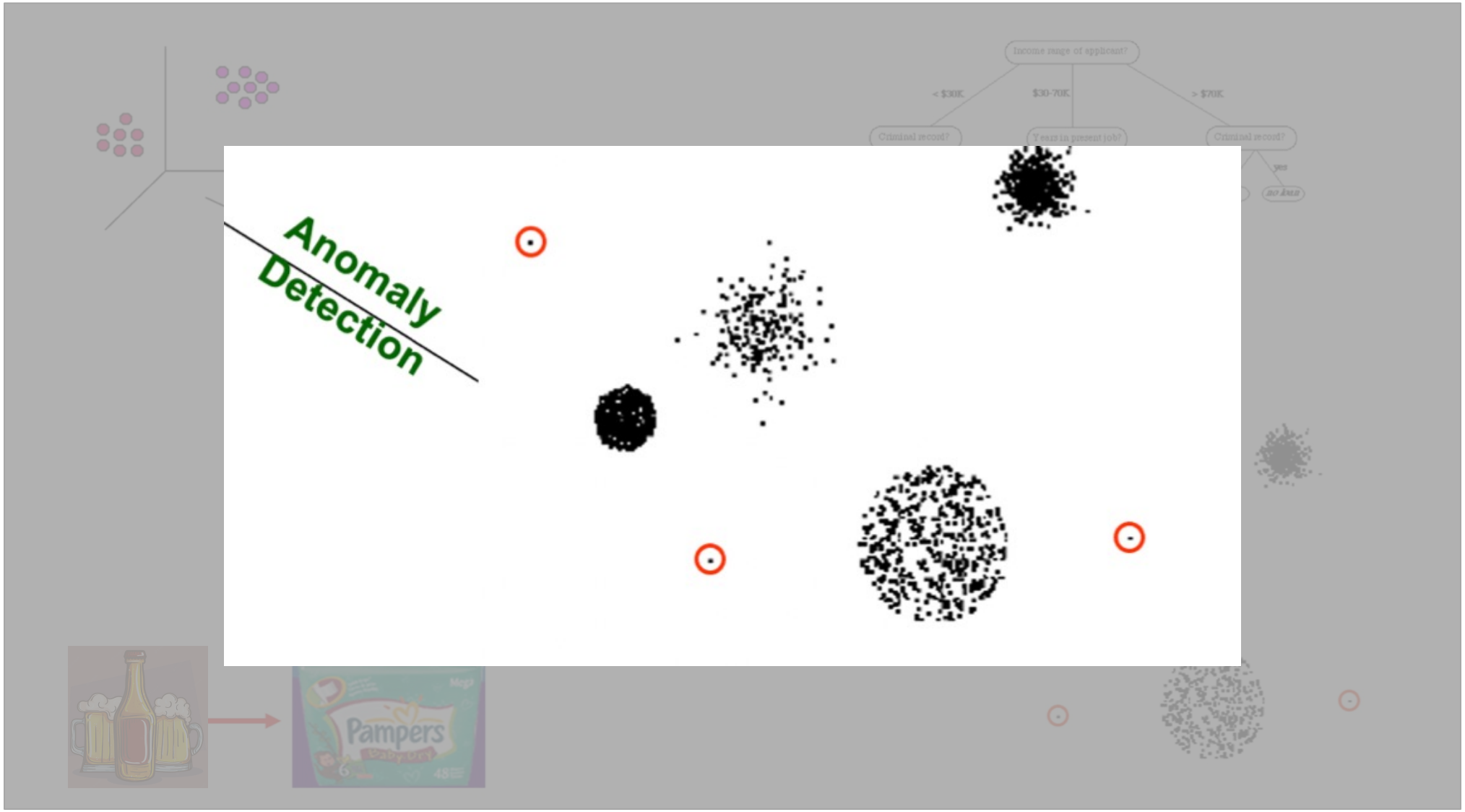
Predictive Modeling

Association Rules

Anomaly Detection

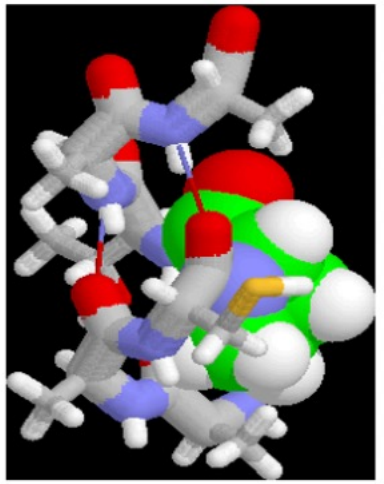


Data Mining



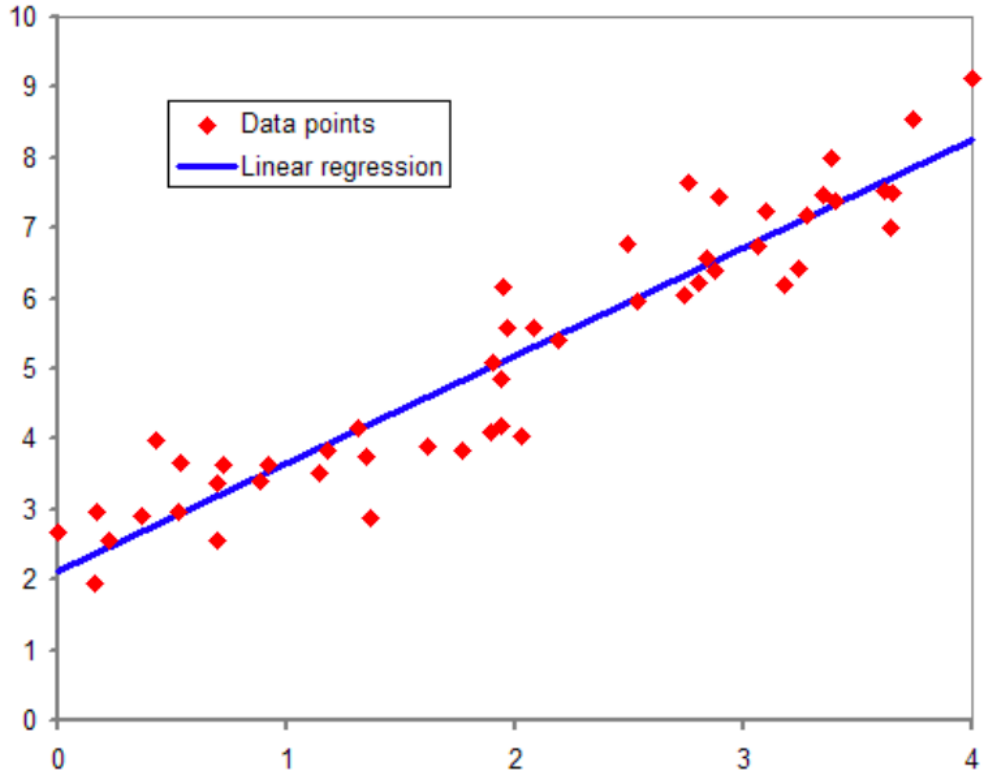
Introduction - Classification

- Classifying credit card transactions as legitimate or fraudulent.
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data.
- Identifying intruders in the cyberspace.
- Predicting tumor cells as benign or malignant.
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil.



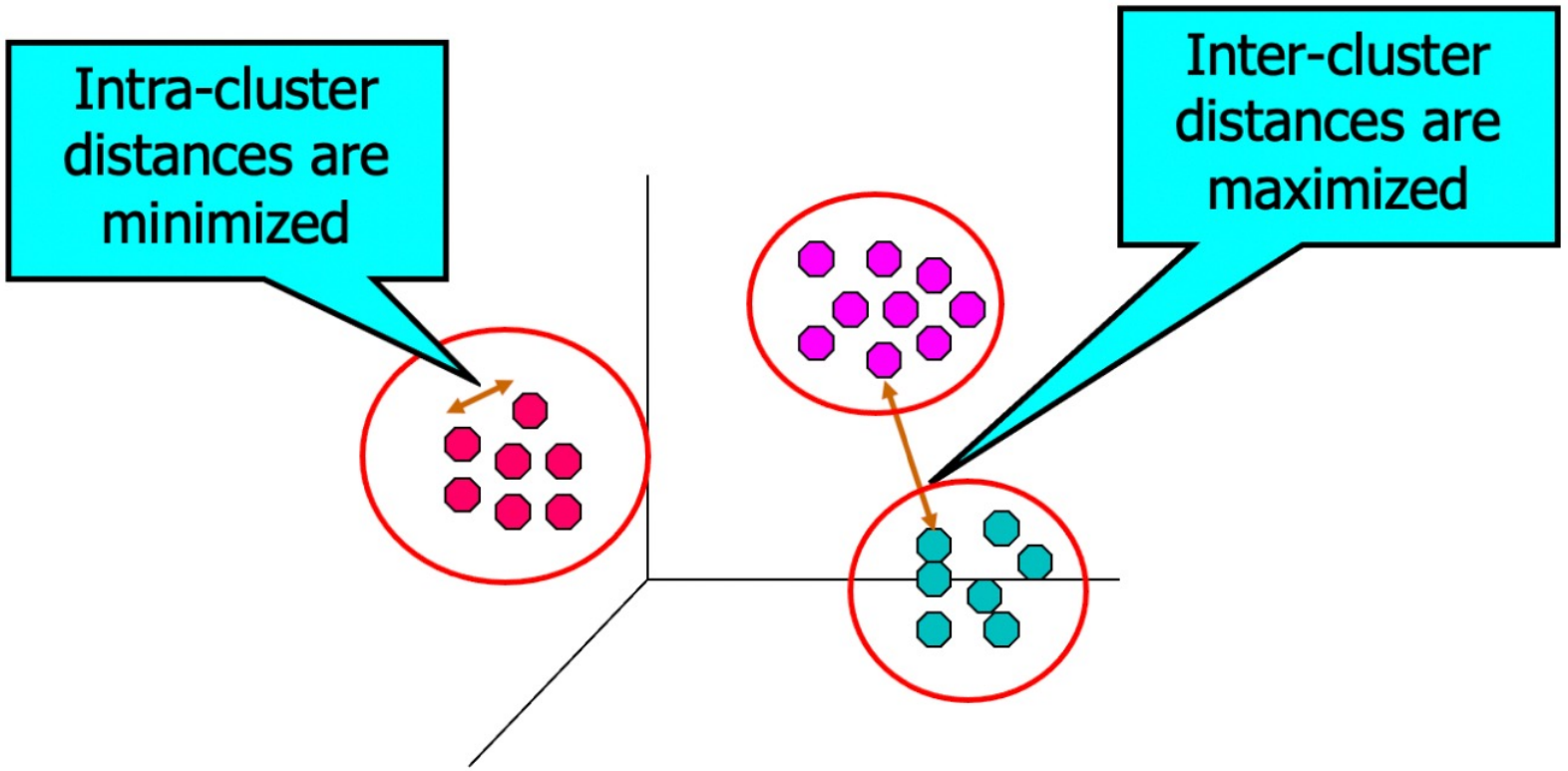
Introduction - Regression

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



Introduction - Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Introduction - Clustering

- Understanding
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- Summarization
  - Reduce the size of large data sets

Introduction - 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.

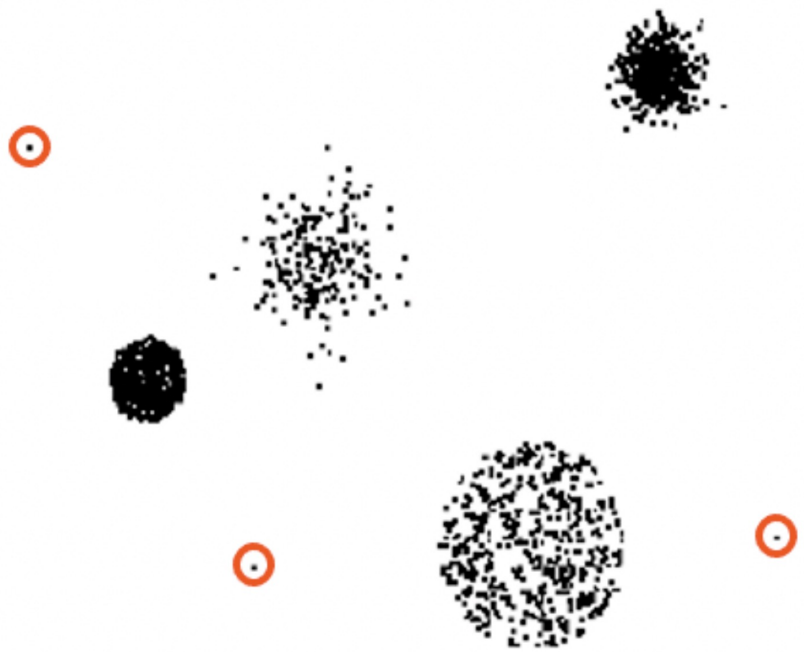
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

- Implication means co-occurrence, not causality!

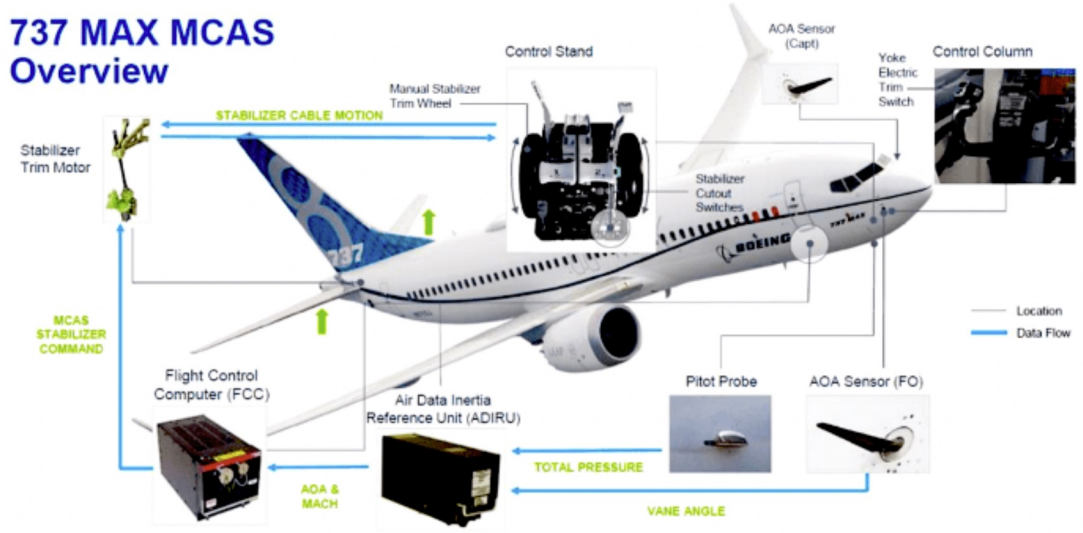
Introduction - Anomaly Detection/Outlier Detection

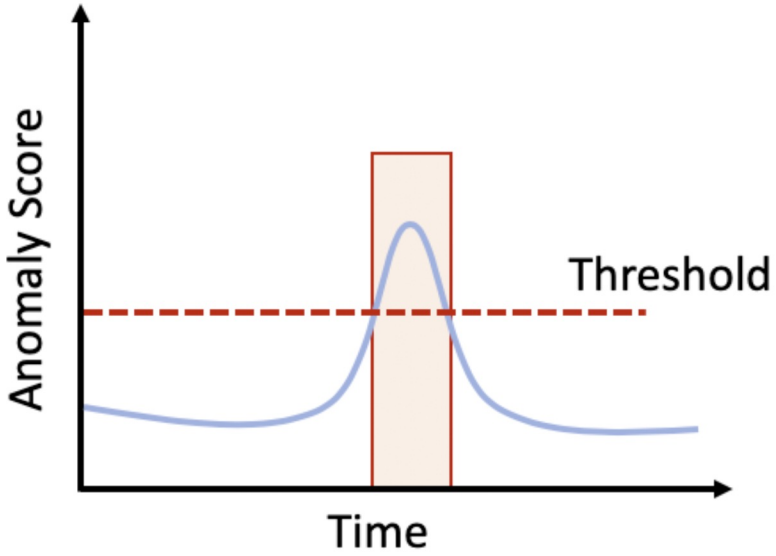
- Detect significant deviations from normal behavior.



Introduction - Anomaly Detection/Outlier Detection

- Detect significant deviations from normal behavior.





Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis
  - hypothesis generation and evaluation