DAT550 Data Mining

Introduction

Structure of the Course

- 2 + 2 + 2 = 4 hours lecture + 2 hours lab
- Sometimes hands on in the class either paper based or with laptop depends if there is time
- 3 mandatory assignments (pass/fail) all of them must be approved to be eligible to take the exam
- A mini project (counts for 40% of the grade).
  - The grades are assigned based on a report (around 10 pages) and code.

Exercises and Hands-on

- In class room paper based exercises
  - Will prepare you for the theoretical questions
  - Helpful for exam
- Hands-on with the python notebook
  - Will give you practical perspective on the theory
  - Helpful for the lab assignments and project

Assignments

- Three assignments
- Individual
- Assignment 1: Related to data preprocessing, feature selection and dimensionality reduction
- Assignment 2: Classifier most likely implementation of decision tree from the scratch
- Assignment 3: Clustering most likely LSH implementation
- Must be approved in person by the TA or me

Mini Project

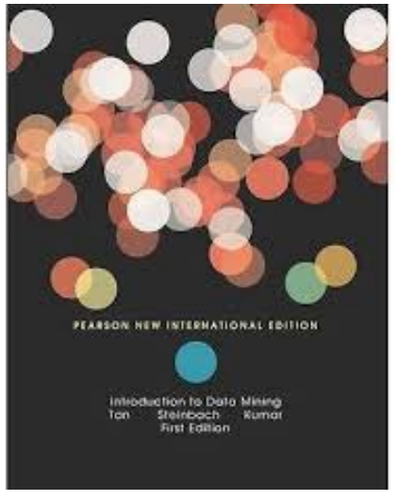
- Work in groups of 3
- You may choose an interesting topic that inspires you
  - Must be approved by me
  - Start looking already
  - Could be using any library or programming language
    - Classification, clustering, deep learning everything is allowed
- If you are unable to find on your own any we will help you or we have some topics
- Higher the quality of the project and the report better grade
  - 60% for the project idea, execution and successful results
  - 40% for the report

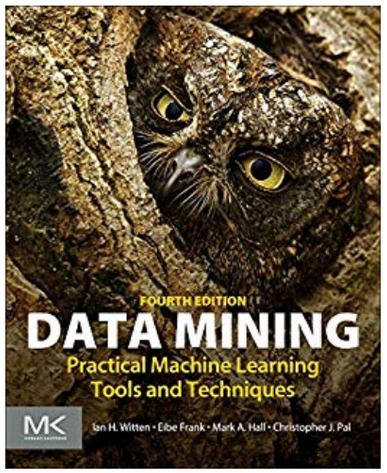
Autograder

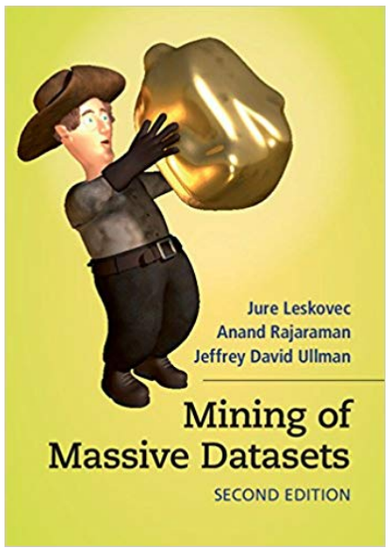
- Register with your with your GitHub account
- Go to ag3.ux.uis.no
- Login with your GitHub account
- Give your full name and valid student id (invalid accounts will not be approved)
-

Syllabus

- Introduction to Data Mining: Tan, Steinbach, Kumar
  - All chapters and appendices except chapter 9
- Data Mining Practical Machine Learning Tools and Techniques, 4th Edition, by Witten, Frank, Hall and Pal
  - Chapter 10 (Deep learning)
- Mining Massive Datasets (mmds.org)
  - Chapter 3 (Locality Sensitive Hashing)
- All slides, exercises, hands-on exercises, lab assignments and other additional reading material provided during the lecture







What is Data?

Attributes

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**

Objects

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

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

What is Data Mining?

"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 Machine Learning?

"The field of study that gives computers the ability to learn without being explicitly programmed"

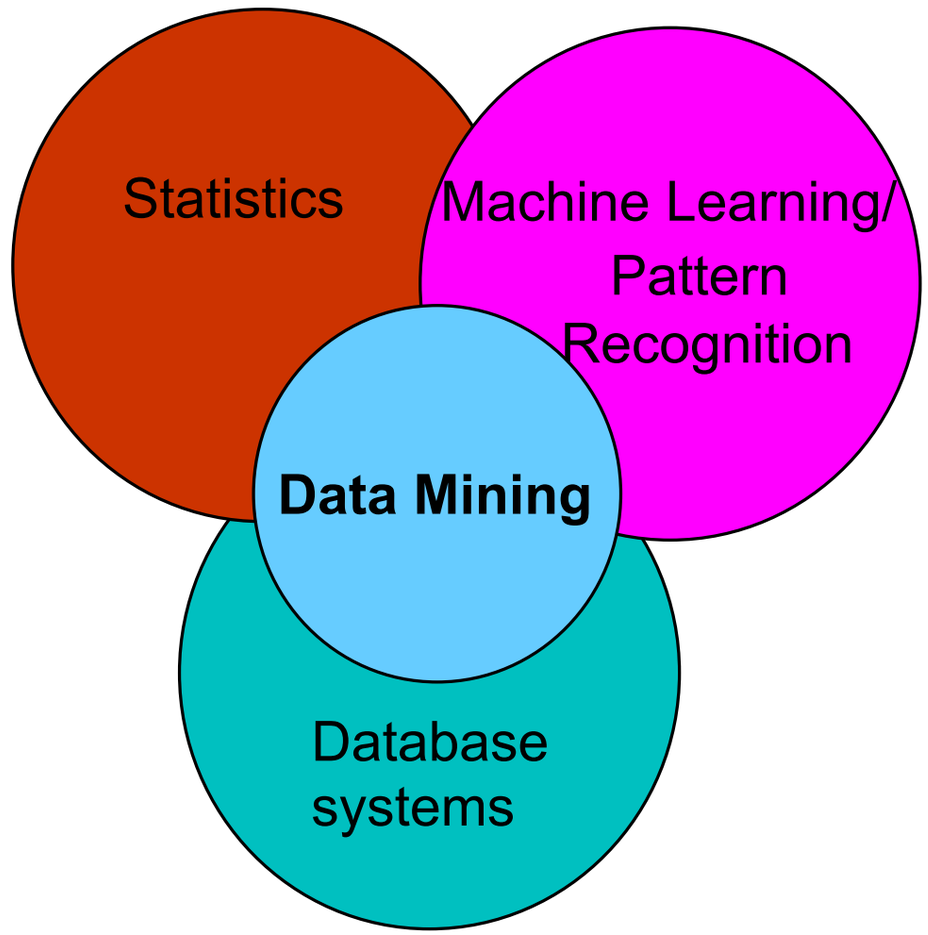
-Arthur Samuel

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

-Tom Mitchell

Data Mining vs Machine Learning

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems



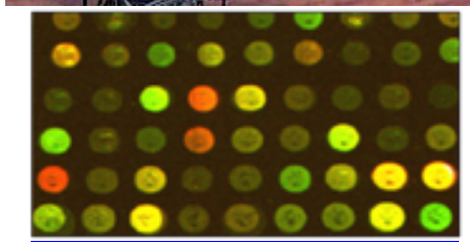
Why Mine Data? Commercial Viewpoint

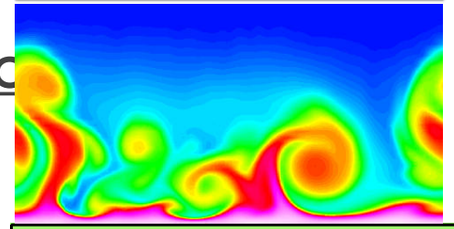
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at supermarkets
  - Bank/Credit Card transactions

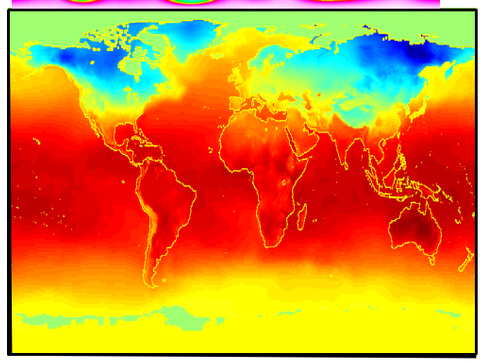
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)
  - Attract new customers
  - Retain customers

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (TB/hour)
  - Telescope data scanning the universe
  - Detecting galaxies, red giants, exoplanets etc
  - LHC (Large Hydron Collider)
    - Collects Peta byte per second (http://www.sixtysymbols.co petabyte_LHC.htm)
    - Detecting new particles
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

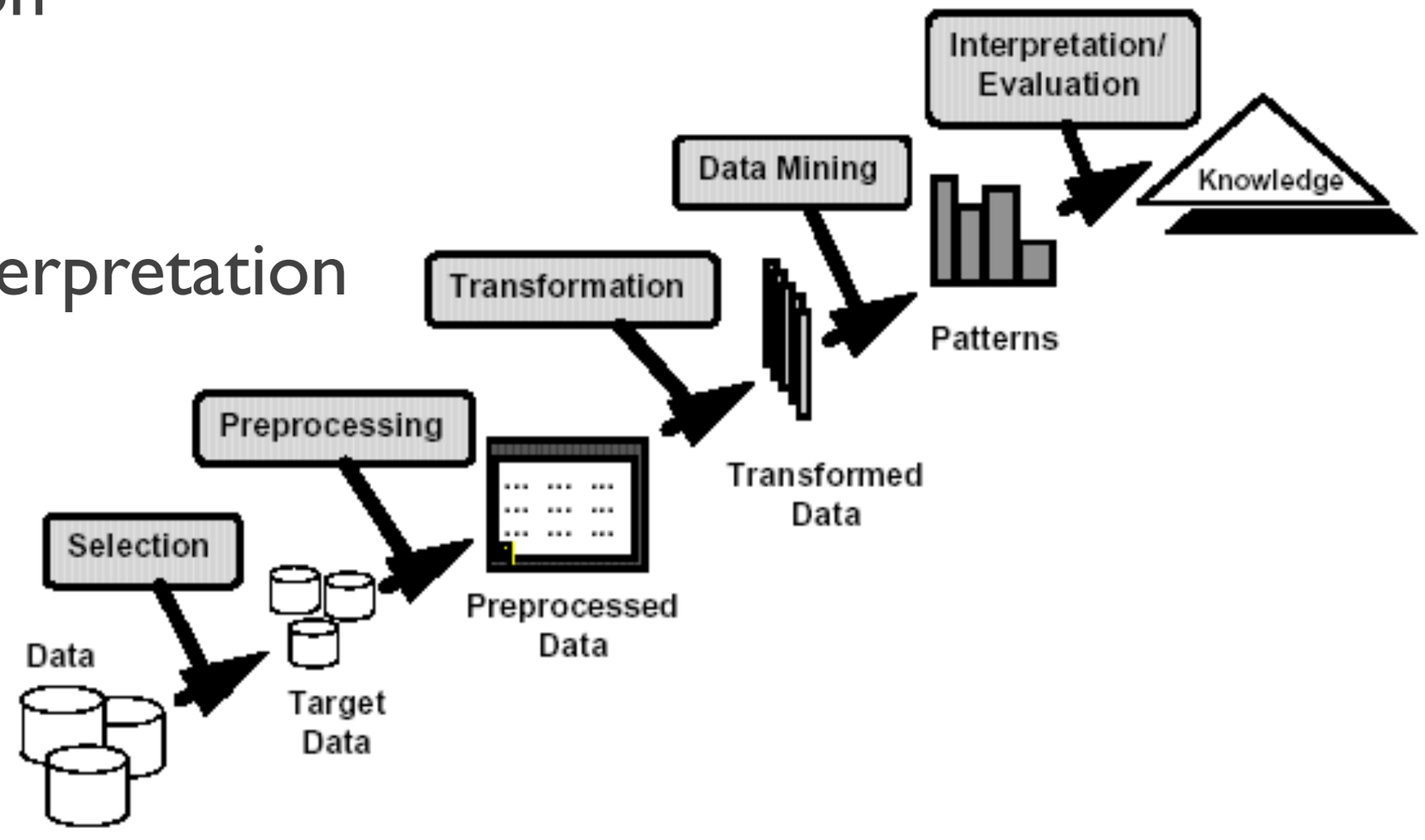






KDD Process

- Data selection
- Preprocessing
- Transformation
- Data Mining
- Evaluation/Interpretation



Data Mining Tasks

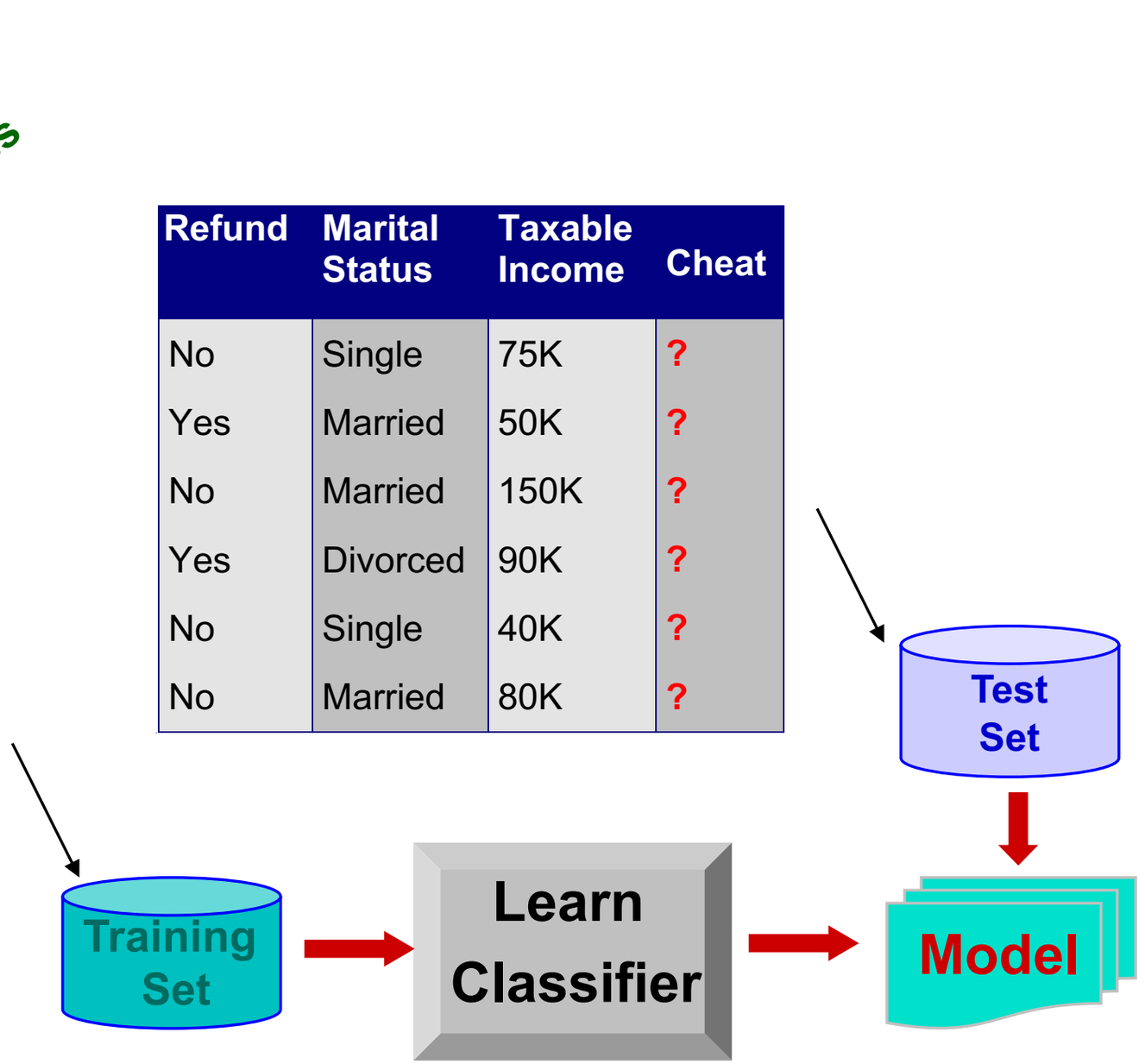
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
    - Supervised
    - Unsupervised
    - Semi-supervised
- Descriptive Methods
  - Find human-interpretable patterns that describe the data.
    - Rule Mining
    - Frequent patterns

Classification Example

categorical | categorical | continuous | class

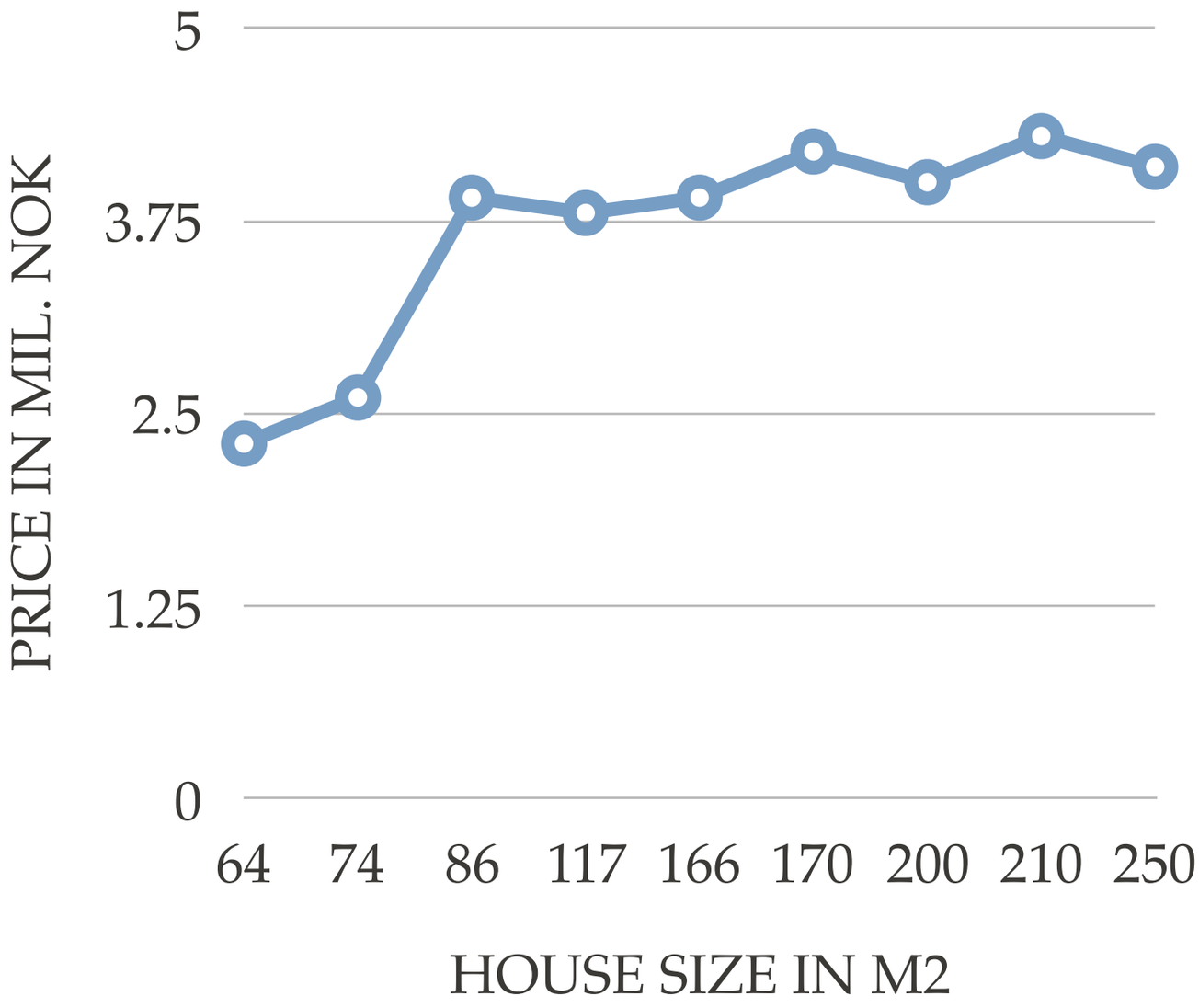
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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



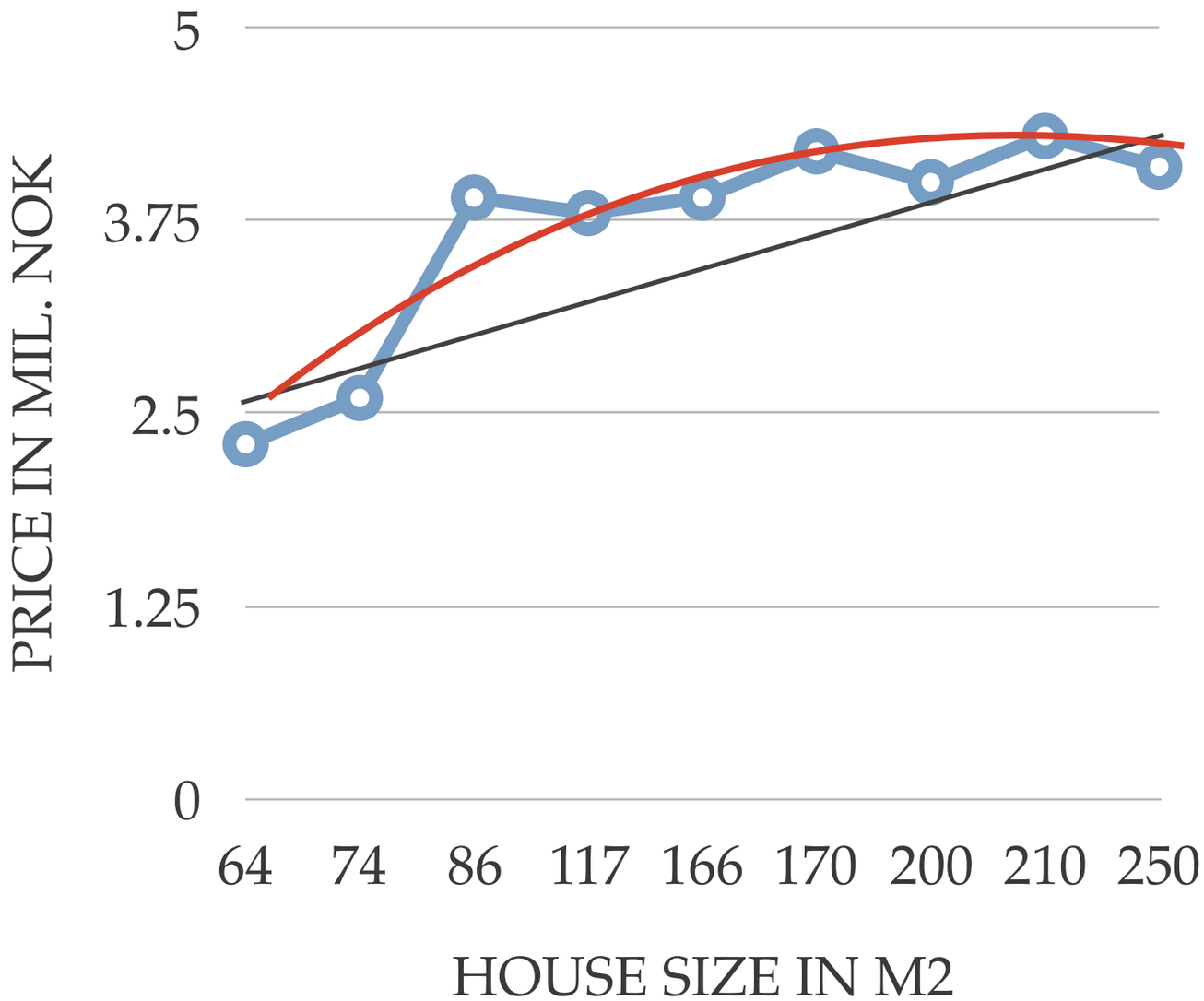
Supervised Learning

Given this data, a friend has a house 80 square meters - how much can they be expected to get?



Supervised Learning

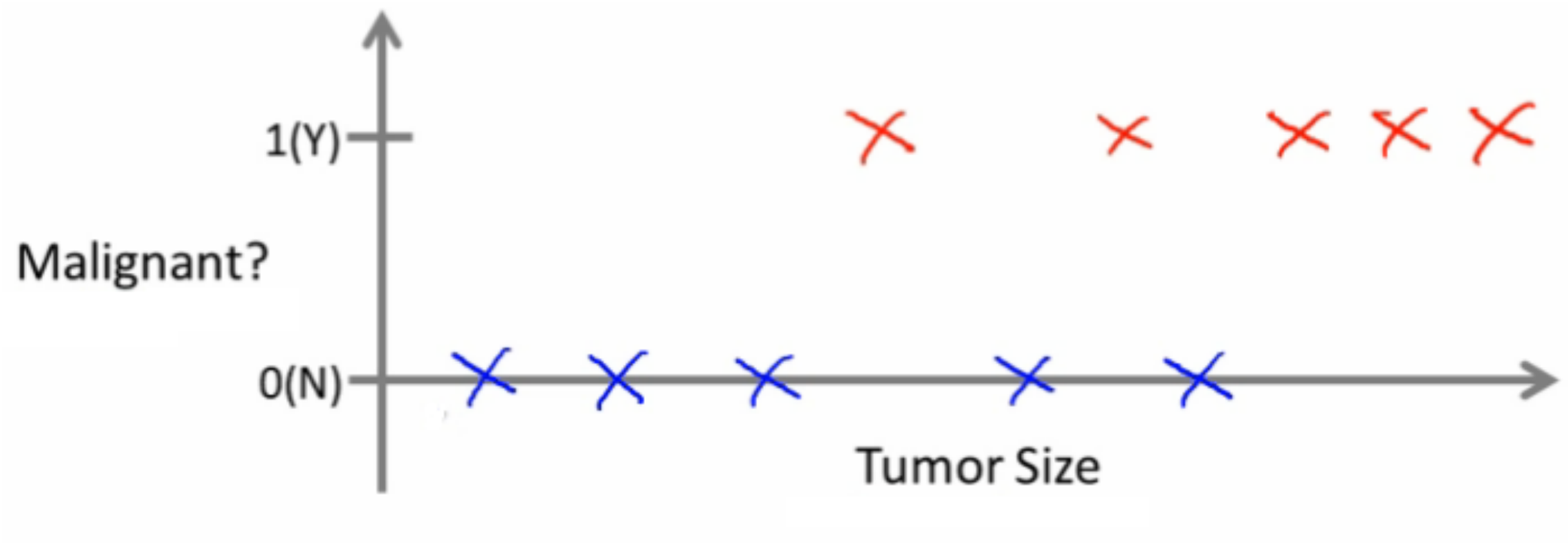
- Straight line through data
  - May be 2.6 Mil. NOK
- Second order polynomial
  - May be 3 Mil. NOK
- Each of these approaches represent a way of doing supervised learning



What is Regression?

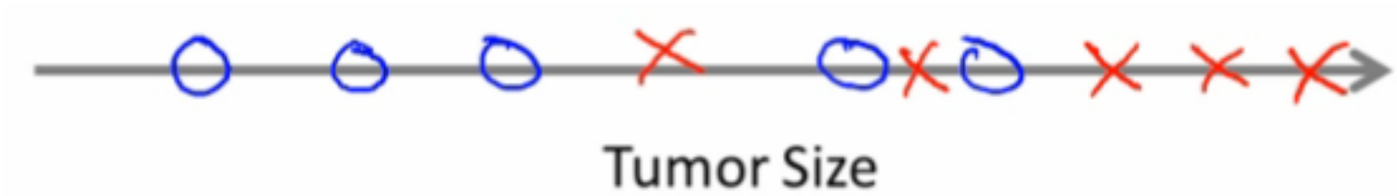
- Given a set of data points/instances along with "correct" answers or labels (training data)
- Feed it to an algorithm which can learn a model and predict the correct value for an unseen data point (test data)
- The learned model is an approximate representation of the training data
- Predict continuous valued output (price in the previous example)
- The algorithm should produce more right answers (goal)

Another example



- Given 10 examples
- 5 positive, 5 negative
- Can you estimate prognosis based on tumor size?

Classification Example

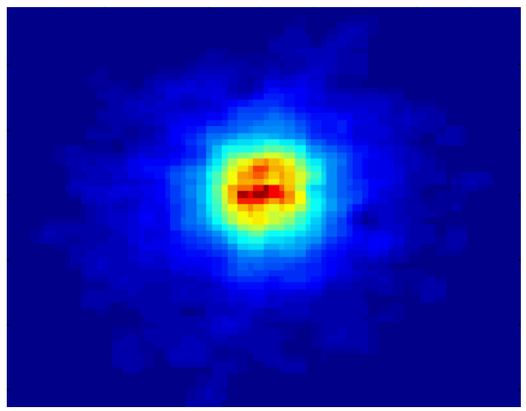


- It could also be multiple classes
  - 0 - benign
  - 1 - type 1
  - 2 - type 2
  - 3 - type 4

Another Classification Example

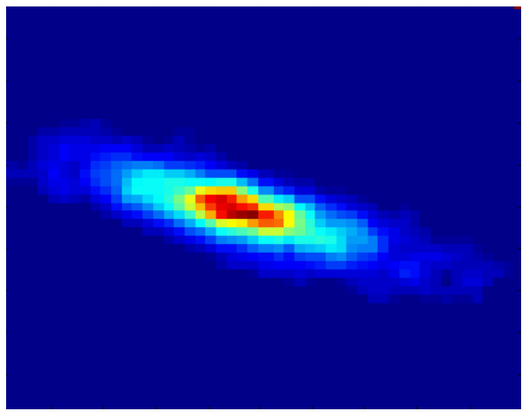
Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Early

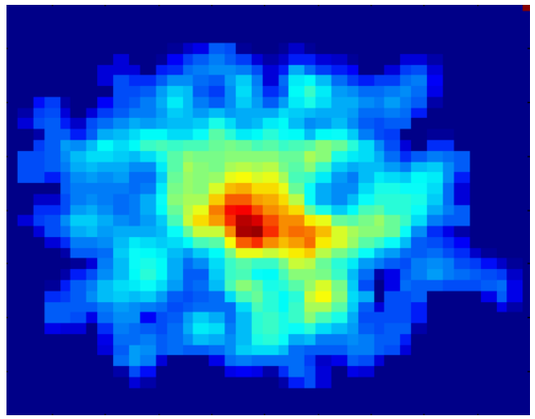


Class:
- **Stages of Formation**

Intermediate

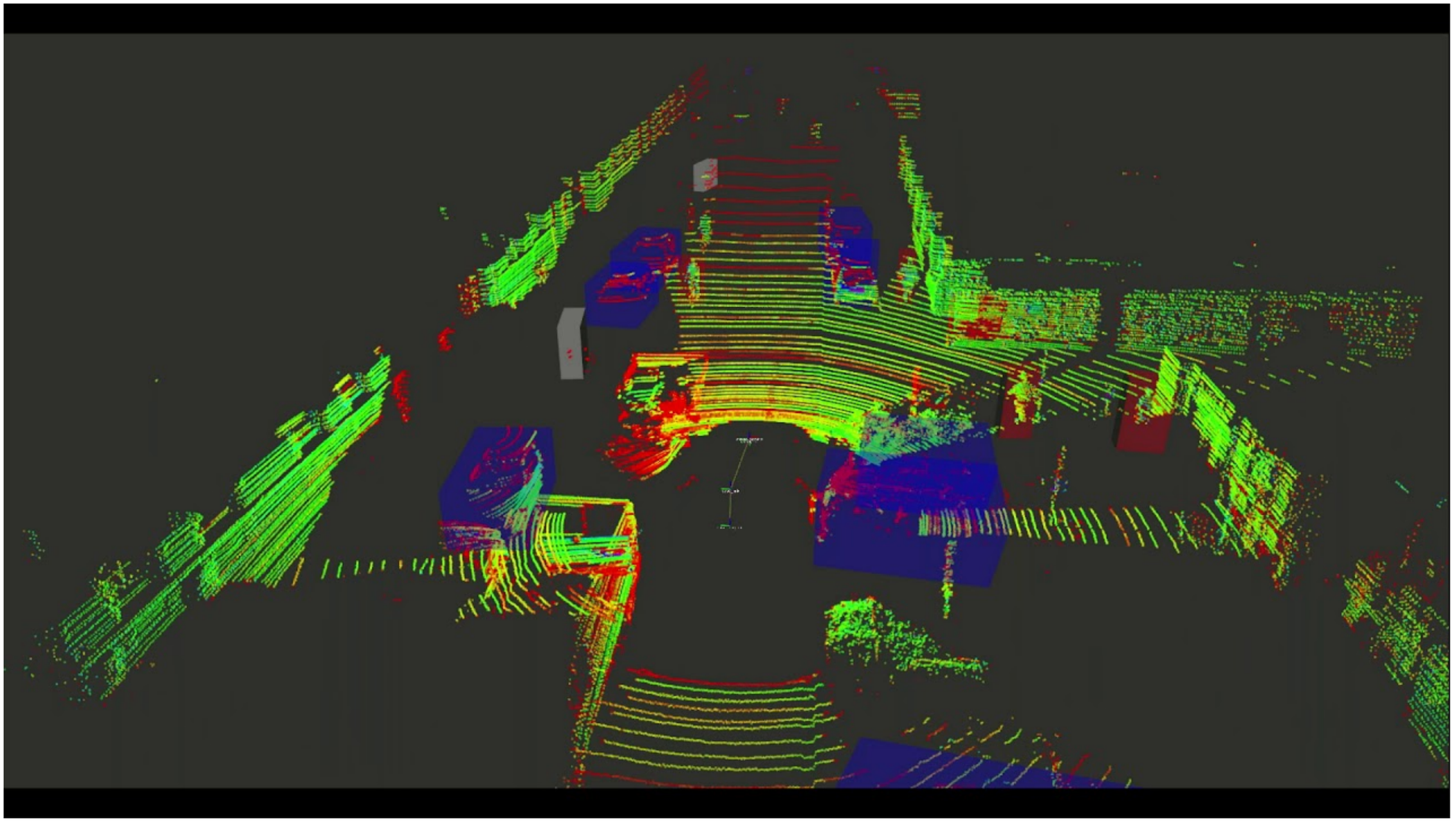


Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Self-driving Cars



Multi-label Classification Example

- For example,
  - A researcher publishes in machine learning conferences and databases conferences
  - Classifying researcher profile leads to multiple labels
- Fact classification
  - Text classification
  - A fact can be both positive and false or negative and true etc

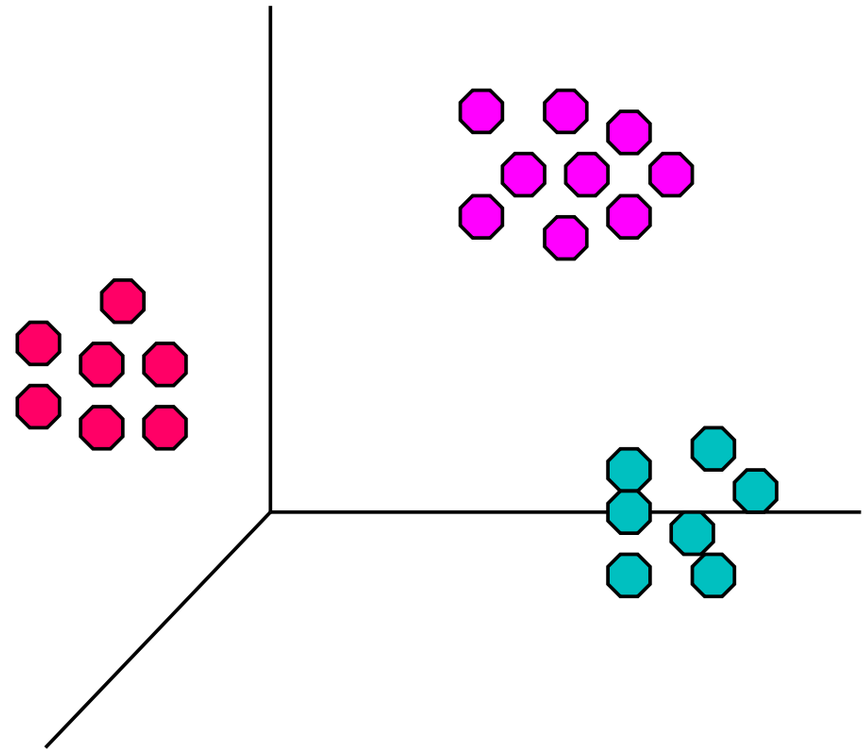
Unsupervised Learning

- In unsupervised learning, we get unlabeled data
- One way of doing this would be to cluster data into to groups
- Group data points which are similar together, while separating dissimilar items as much as possible
- This is a clustering algorithm

Clustering Example

Intracluster distances are minimized

Intercluster distances are maximized



Clustering Example

Google News

Headlines

More Headlines

Trump's idea to declare national emergency raises legality questions | TheHill

The Hill • one hour ago

- **White House signals some compromise in ending U.S. government shutdown**

  AOL • 3 hours ago

- **White House: President Trump means it when he says the government could be shut down for months or y**

  Fox News • 5 hours ago

- **Democrats must prove they are worthy of their House majority | TheHill**

2 Americans Said to Have Joined ISIS Caught on the Front Lines in Syria

The New York Times • one hour ago

- **Syria conflict: Bolton says US withdrawal is conditional**

  BBC News • today

View more

A very simple news clustering example

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering: Market Segmentation

- Market Segmentation:
- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Association Rule Mining

- 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}

What can we do with the rules discovered?

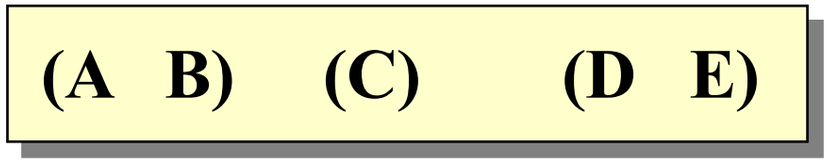
- Marketing and Sales Promotion:
  - Get the rule discovered be
  - **{Bagels, … } --> {Potato Chips}**
- Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
- Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
- Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rules Application

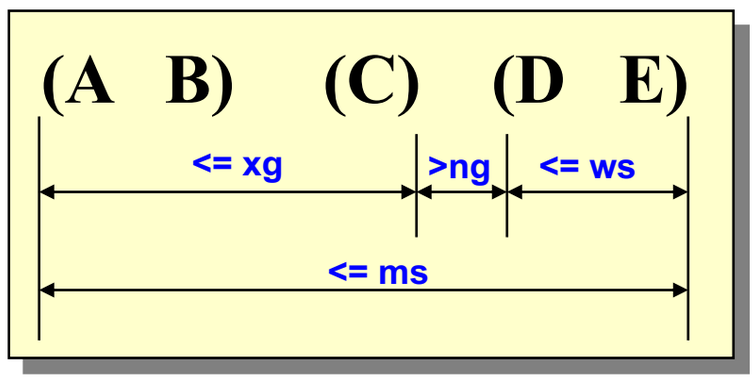
- Inventory Management:
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.

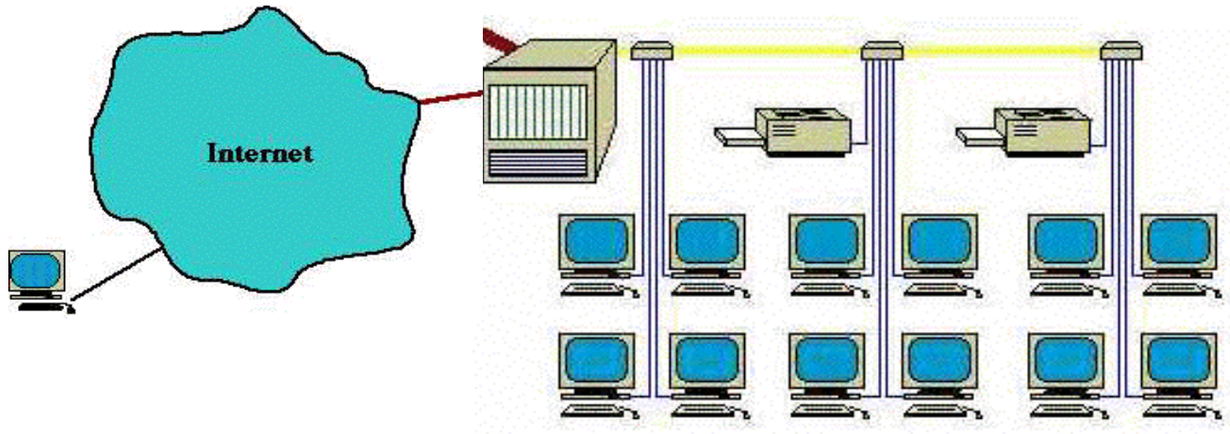


- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Spam detection Twitter
  - Fake news detection



Challenges in Data Mining

- Scalability
  - Terabytes, petabytes of data is common
  - Some solutions: Sampling, distributed processing
- Dimensionality
  - Self-driving cars have to deal with data having thousands of dimensions
  - Gene sequence data
  - Solution: Dimensionality reduction
- Heterogenous or complex data
  - Dealing with cross-domain data
  - Categorical, sequential, continuous data

Data Mining Challenges

- Data ownership and distribution
- Privacy concern
- Data quality
- Evolving/streaming data

Literature

- Tan, Steinbach and Kumar, Chapter 1