COSC2110/COSC2111
Data Mining

Vic Ciesielski
Department of Computer Science
RMIT
vic.ciesielski@rmit.edu.au
14.08.16
I'll be at most of the labs

ABOUT ME

1. Researcher in Machine Learning and Evolutionary Computing

2. Data mining for REALas.com

When	
Feb-2011	3 investors want a recommender system
Mar-2011	Start development, data collection
Jun-2011	Discover prices can be predicted with surprising accuracy
Jul-2011	Switch to developing prediction system
Oct-2012	Live running, "That's my Offer"
Jun-2014-Nov-2014	Major Redevelopment
Oct-2016	ANZ buys REALas
Now	12 People employed

PRACTICAL WORK/ASSIGNMENTS

Standard Assignments
OR
Kaggle Competition
OR
Deep Learning
AND
5 minute talk on ethical considerations

```
 88 M Information Technology
 20 Master of Data Science
  8 M Computer Science
  7 Master of Analytics
  7 M Stats & Operations Research
  6 Exchange Inbound Students UGRD
  5 BEng(Comp&NetEng)(Hons)/BCompS
  3 BEng(TelecomEng)(Hons)BCompSc
  2 Bachelor of Analytics(Hons)
  1 Program Name
  1 Exchange Inbound Students PGRD
  1 B Computer Science (Hons)
```

MATH2319 Machine Learning
COSC2670 Practical Data Science

TOOLS FOR DATA MINING

	WEKA	R	PYTHON
Algor-ithms	Widest Range	Most of the popular ones	Most of the popular ones
Progr-amming	None	Programming in R language	Python
Data Prep	Wide range of filters	Some support	Some support

- Familiarity with Unix scripting is invaluable
- For any task relating to transforming data there is very likely a Unix program that does it.

INFORMATION ABOUT THE COURSE

- Lectures: Vic
- Tutorials and Labs: Kendall Taylor and Abhinay Kathuria
- Topics
  - Introduction
  - Classification
  - Clustering
  - Association Finding
  - Attribute Selection
  - Visualisation
  - Unix Tools for Data Manipulation
  - Neural Networks
  - Text Mining
  - Evolutionary Classifiers
  - Case Study
  - Special Topics
  - Assignment 1 due at the end of week 6
  - Assignment 2 due at the end of week 11
- Course Guide
- Course will use Canvas
- Course will use Lectopia/echo
- Plagiarism Warning

SUMMARY TO FIRST ASSIGNMENT

1. The process of knowledge discovery in data bases
2. Data Mining Tasks
   - Classification
     - ZeroR
     - OneR
     - Nearest Neighbour
     - Decision Tree
     - Error Estimation
     - Neural Network [Assignment 2]
   - Numeric Prediction, time series prediction
     - OneR
     - M5P
     - Nearest Neighbour
     - Neural Network [Assignment 2]
   - Association Finding
     - Apriori
   - Clustering
     - K-Means
     - EM
   - Attribute Selection

WHY IS DATA MINING NECESSARY

Data comes like water out of a fire hydrant. You can't drink it (Anon)

We are drowning in information but starving for knowledge (John Naisbett)

- Hardware advances in data collection and storage have far outpaced software advances in data analysis and manipulation.
- Organizations collect more data than they can handle.
- Data that may never be analysed is still collected out of fear of missing something that may be important.
- As databases grow decision making directly from their contents is not feasible; knowledge derived from the data is needed
- Supermarket chains, credit card companies, banks routinely generate gigabytes of data daily
- Exponential growth of web pages
- Entering the era of ~~tera~~ and peta and exa bytes
- Social Media: Google, Facebook, Twitter
- FANG (Facebook, Amazon, Netflix, Google)

KNOWLEDGE DISCOVERY & DATA MINING

1. *Knowledge Discovery in data bases (KDD)* is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad]

2. 'Golden Nuggets'

3. *Data mining* is the part of the KDD process relating to methods for extracting patterns from data [Fayyad]

4. *Data mining* is a problem solving methodology that finds a logical or mathematical description, of a complex nature, of patterns and regularities in a set of data [Decker and Focardi]

5. In some current usage KDD = Data Mining

6. Knowledge extraction, information discovery, information harvesting, data archeology, data dredging, data pattern processing

7. Emerging new names: 'Data Analytics', 'Predictive Analytics', 'Big Data', 'Data Science'.

EXAMPLES OF NUGGETS

- Fraudulent credit card transactions
- Good/bad loan risks
- New class of stars
- Put beer and disposable nappies together and you'll sell more of each

  http://www.theregister.co.uk/2006/08/15/beer_diapers/

  http://dssresources.com/faq/index.php?action=artikel&id=41
- Put purfume and greeting cards together and you'll sell more of each
- People who bought this product/book also bought these other products/books
- Recognition of specific market segments that respond to specific characteristics
- Telephone customers who will switch to another provider
- Fraudulent international telephone calls
- Ineffective advertising
- Predict pregnancy based on purchasing behaviour, send coupons

  http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?

  pagewanted=all&_r=0

- Dating Site: both men and women's actual behavior differs significantly from their stated tastes and preferences which they outline when they first sign up. In other words, people are not as fussy about partners as they make out.

  http://www.technologyreview.com/view/524081/data-mining-reveals -the-surprising-behavior-of-users-of-dating-websites/

- What goods should be promoted to this customer?

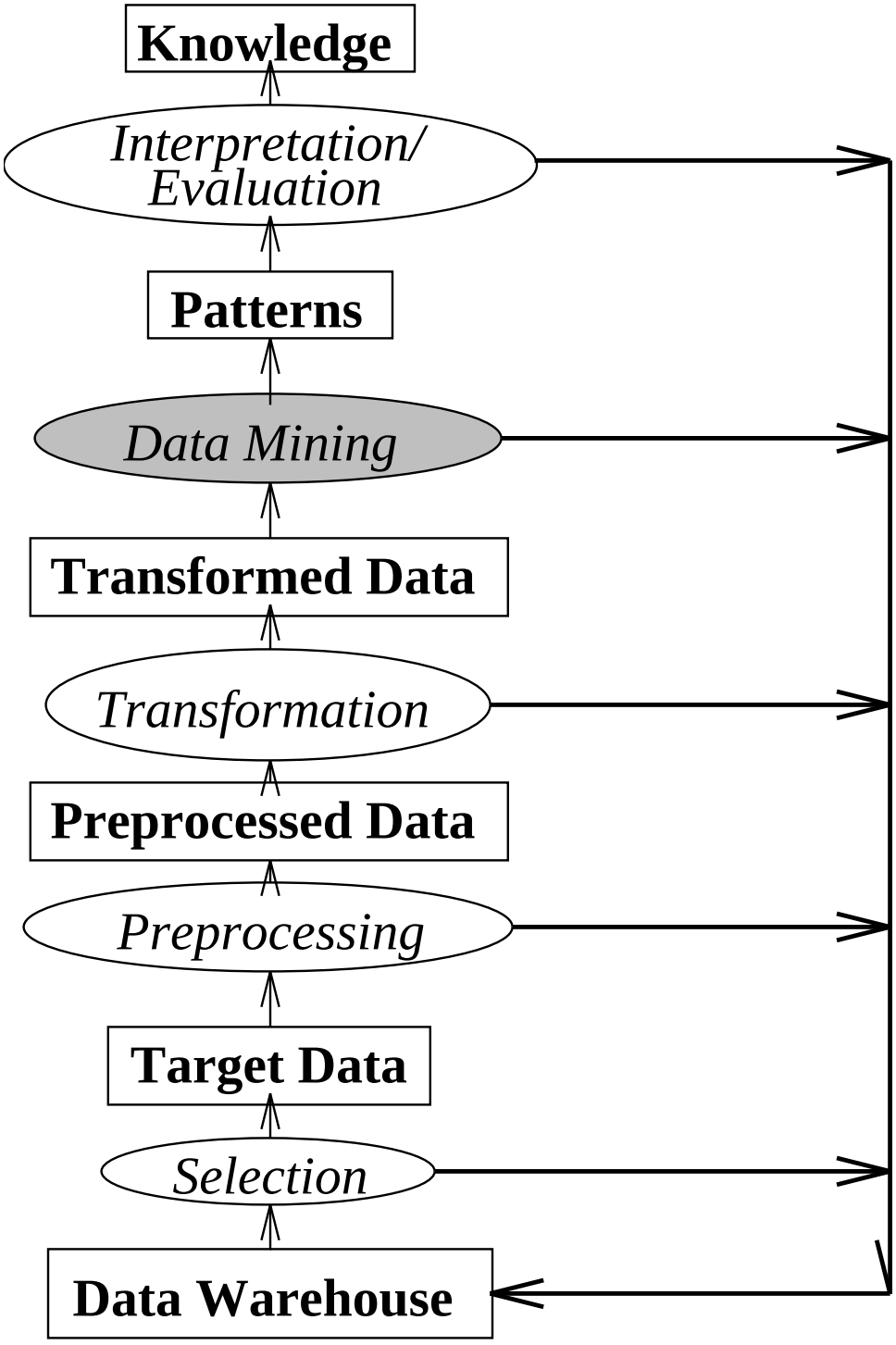
- http://http://www.researchpipeline.com/wordpress/category/data-mining/

- The Economist recently had an interesting article on how insurance companies are increasingly using data mining to "analyze risk." That is, they look through the data which was originally collected for the purpose of better marketing, and use it as a tool to see if you lead an unhealthy life. However, the really interesting point is highlighted by Kashmir Hill, where an exec at a datamining company admits that he's changed his habits because of this. Not his eating habits, mind you. But how he purchases food:

- Just wait. We'll be sending you coupons for things you want before you even know you want them.

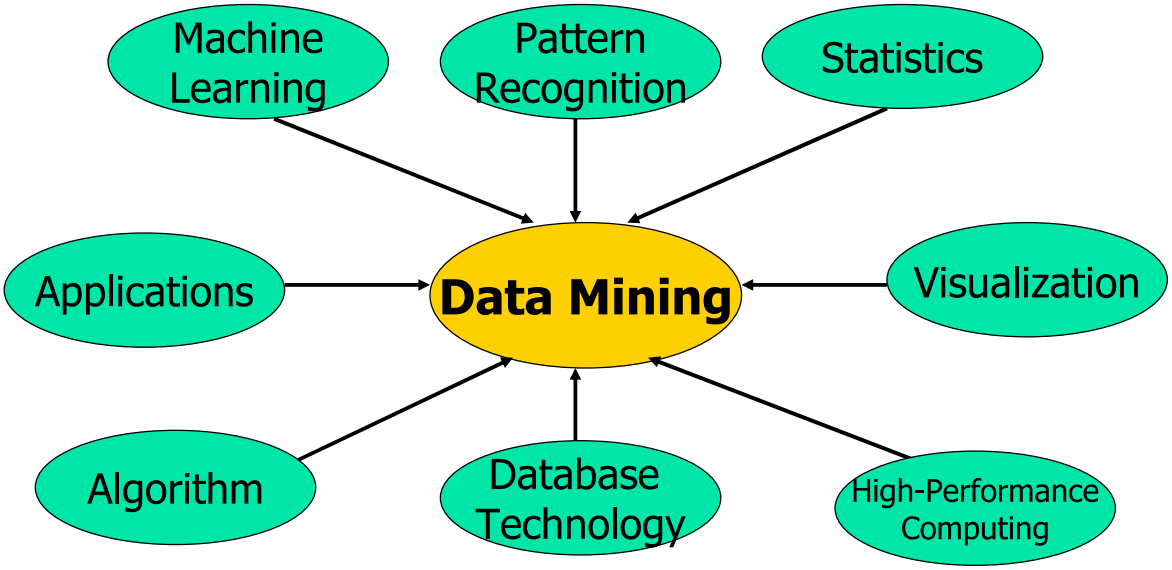
- Marissa Mayer said credit card companies can predict a divorce with 98% accuracy two years before it happens. Considering 50% of marriages end in divorce anyway, that might not be considered impressive.
- Data Mining applications in health care. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf
- http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/
- http://www.dataminingcasestudies.com/
- http://dml.cs.byu.edu/ cgc/docs/mldm_tools/Reading/DMSuccessStories.html
- http://www.newser.com/tag/7595/1/data-mining.html

KNOWLEDGE DISCOVERY IN DATA BASES



INTERDISCIPLINARY FIELD

Data Mining: Confluence of Multiple Disciplines



28

Source: `http://www.cs.uiuc.edu/homes/hanj/bk3/bk3_slidesindex.htm`

Statistics vs machine learning "wars".

EXAMPLE

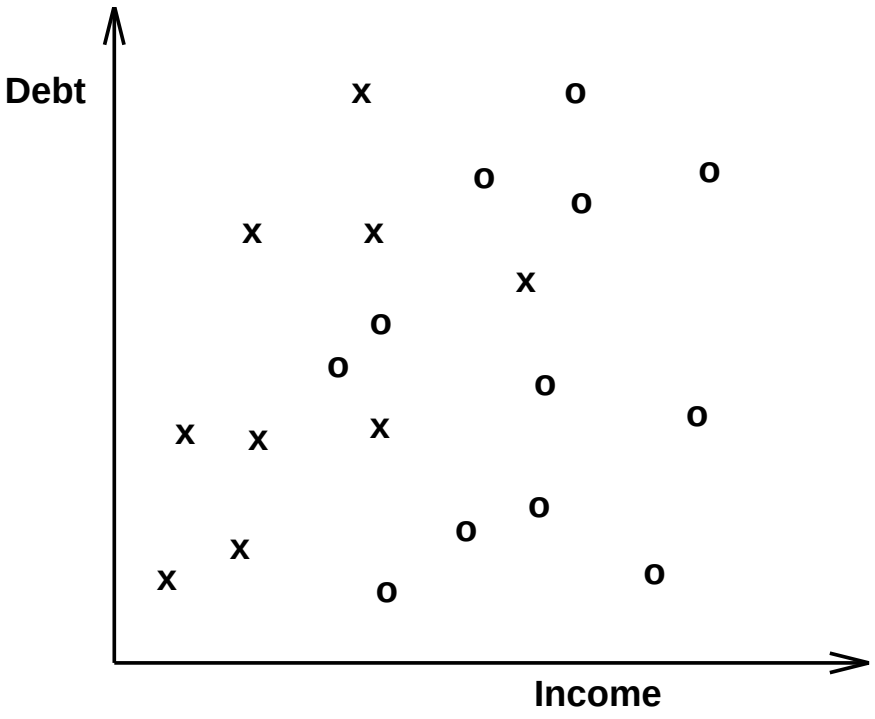
Consider the data from a loans data base:

What can be discovered? Patterns? Regularities?

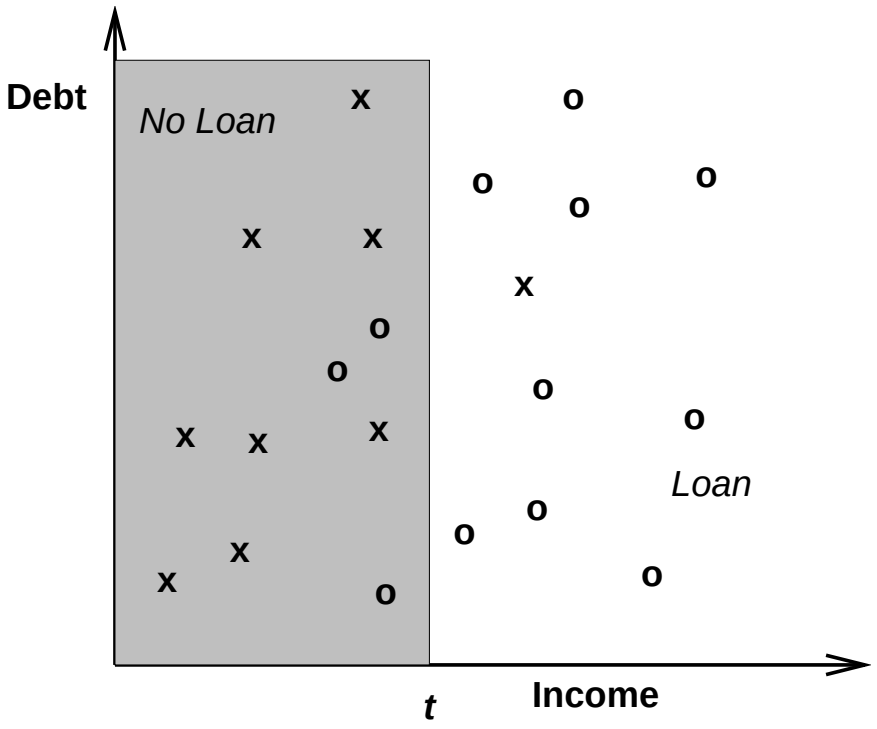
The bank would like to know whether an applicant for a loan will pay it back or not.

Income	Debt	Defaulted?
\$20,000	\$1,000	no (o=good)
\$50,000	\$25,000	yes (x=bad)
.	.	.
.	.	.

SCATTER PLOT



THRESHOLD



Suppose that we 'extract' the classification rule:

if Income $> t$ then grant loan

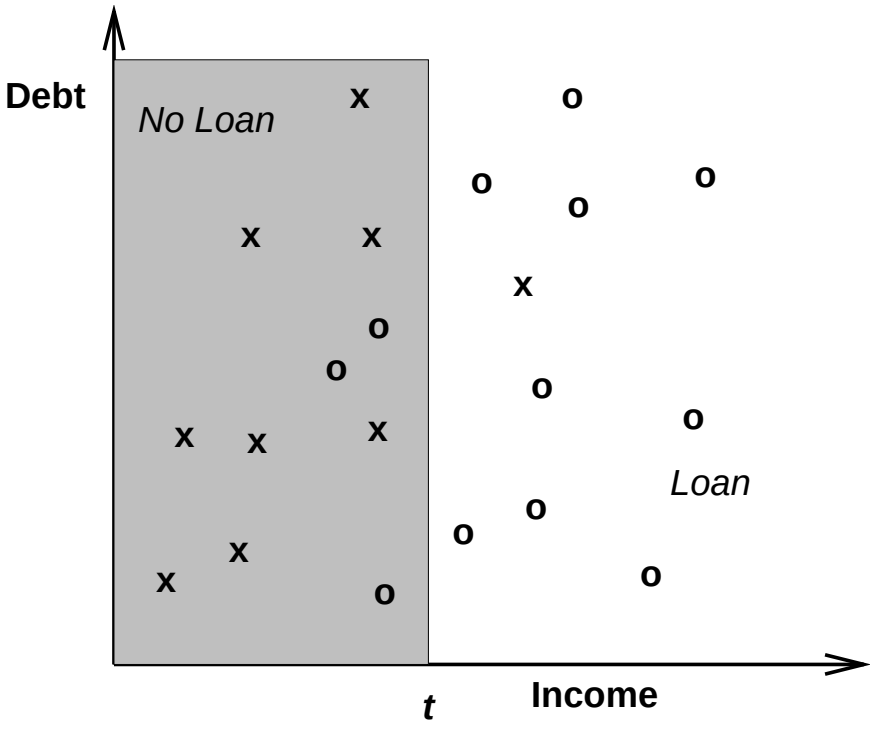
1. How many errors would we make on this data?

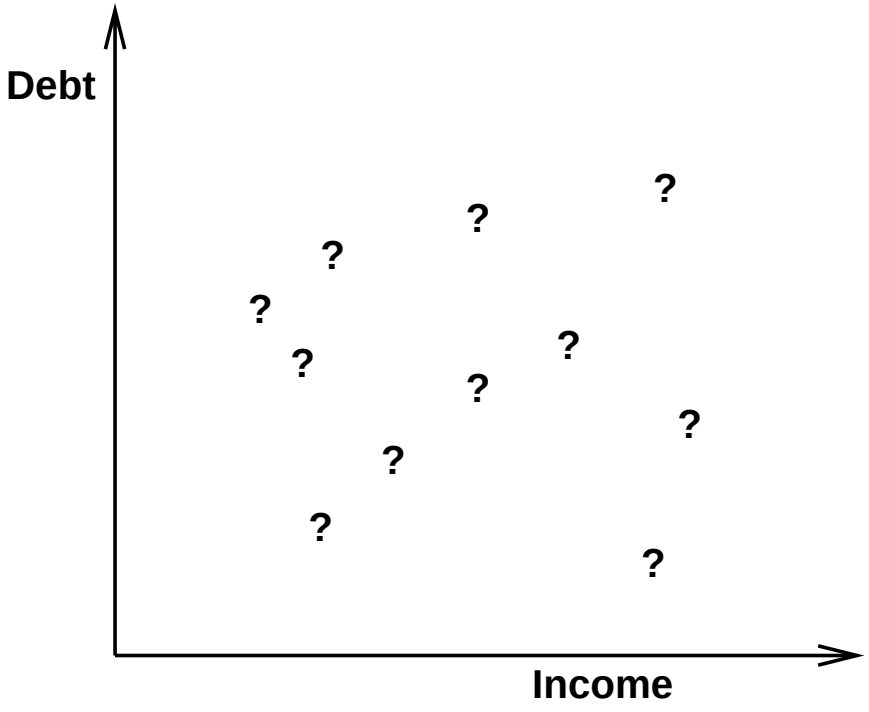
2. How many errors on new customers?

3. How should we choose t

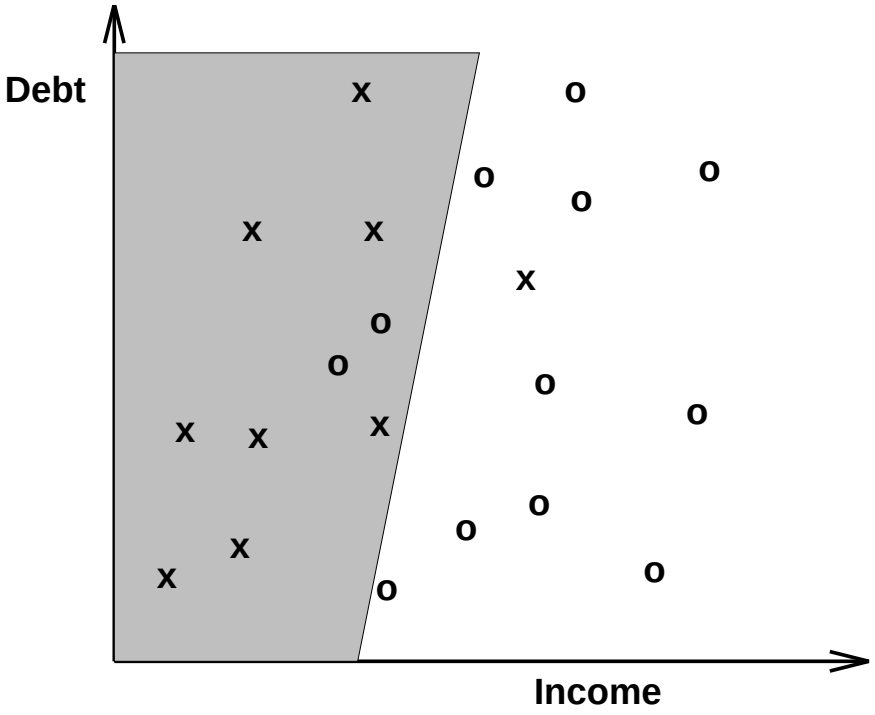
4. Can we find a better rule?

THRESHOLD





LINEAR DECISION BOUNDARY



If the equation of separating line is

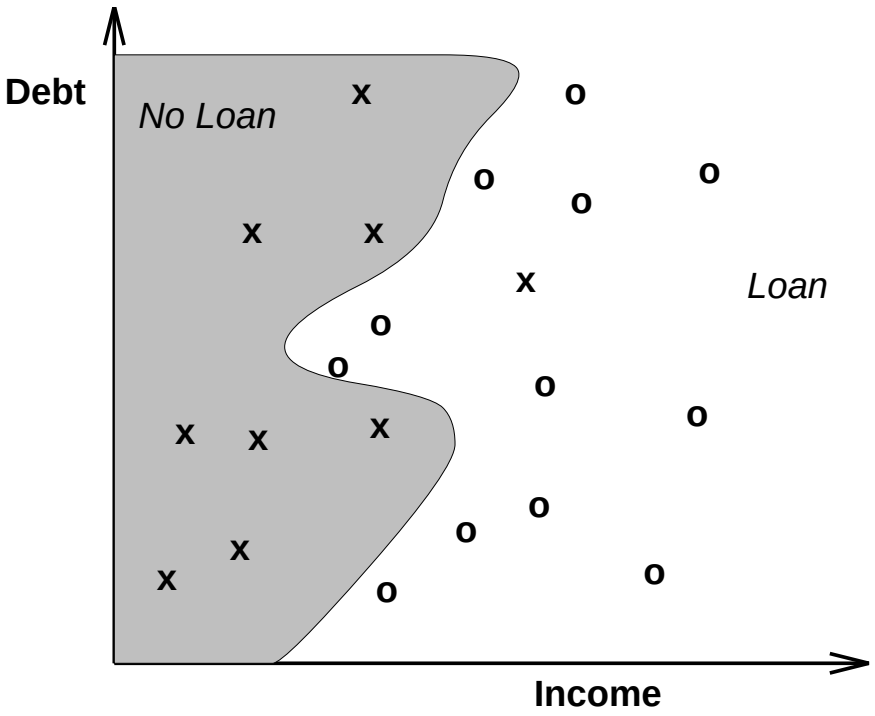
$$Debt = \alpha.income + \beta$$

we can 'extract' the classification rule:

$$if\ Debt < \alpha.income + \beta\ then\ grant\ loan$$

1. How many errors would we make on this data?
2. How many errors on new customers?
3. How can we find the best line?
4. Can we find a better rule?

NON LINEAR REGIONS



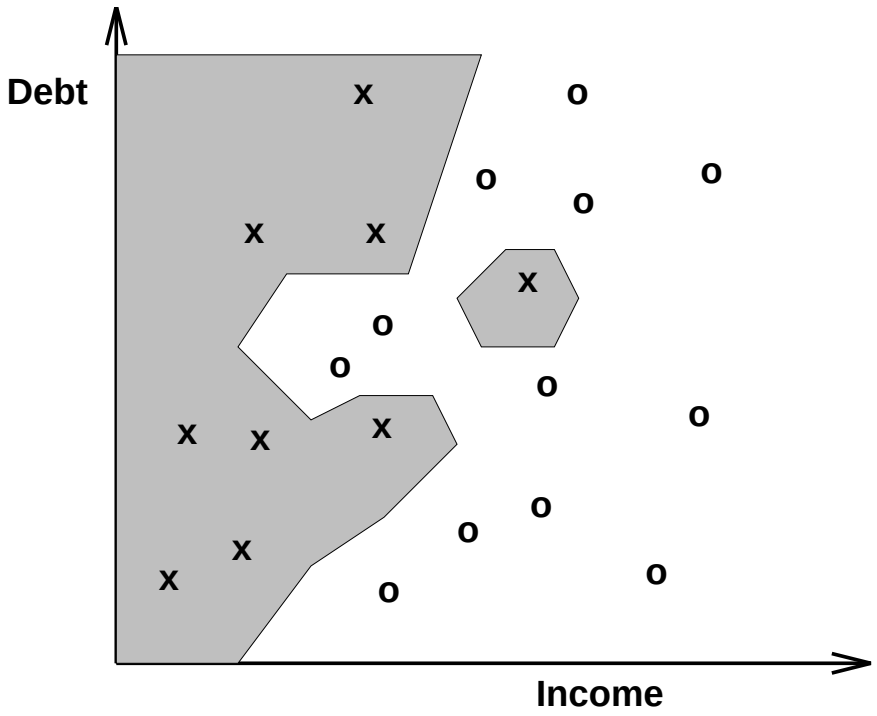
1. How many errors would we make on this data?

2. How many errors on new customers?

3. How can we find the best curve? What is its equation?

4. Many classifiers will give a decision but not the equation of the separating curve.

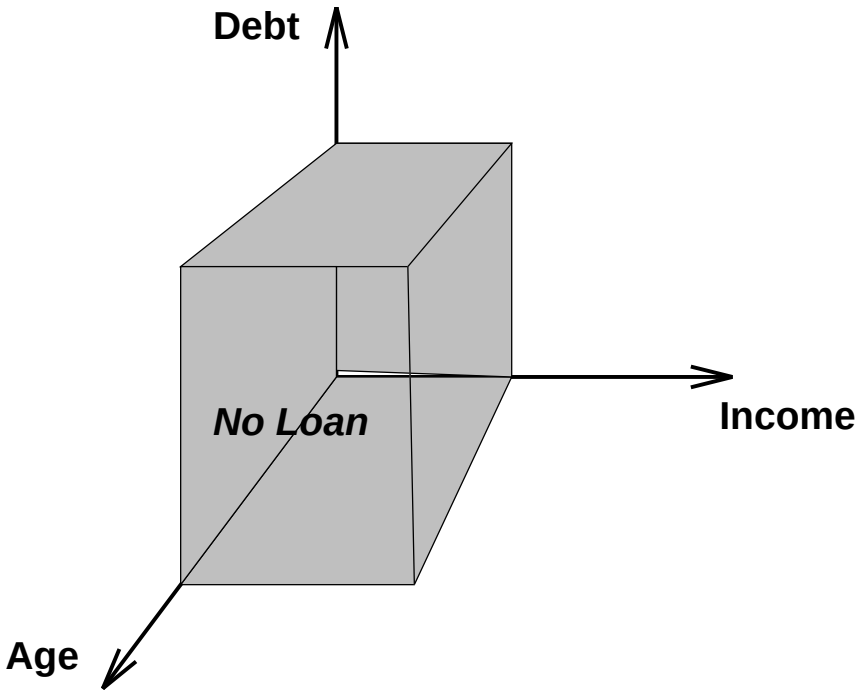
NEAREST NEIGHBOUR



- Each unknown point is given the classification of its closest (k) neighbours
- How many errors on new customers?
- How can we find the best curve? What is its equation?
- Nearest neighbour can give us the regions but not the equation of the separating curve.

HIGHER DIMENSIONS

- Suppose we believe that older people are more likely to pay off loans than younger people. We can include age in the decision.



- If we also include bank balance how can we visualize the result?

- How can we include male vs female and other non numeric data?

CLASSIFIERS

- A *Classifier* is a procedure which takes as input data about a particular case and outputs one of a set of predefined labels for it.

- Example 1: Take as input debt and income, output GRANT LOAN or DONT GRANT LOAN

- Example 2: Take as input patient medical data, output SICK or NOT SICK

- Example 3: Take as input weather data, output WILL RAIN TOMORROW or WILL NOT RAIN TOMORROW

- Example 4: Take as input various attributes of mushrooms, output POISONOUS or EDIBLE

- Example 5: Take as input some measurements of the iris flower, output SETOSA, VIRGINICA, VERSICOLOR

- Note: There can be more than 2 classes

LEARNING OF CLASSIFIERS FROM DATA

- There are literally hundreds of algorithms for doing this
- Many are available in the WEKA package
- Basic classifiers [easy to understand the algorithm]:
  - Nearest Neighbour
  - OneR
  - ZeroR
- A bit more complex
  - Decision Tree
  - Naive Bayes
- Most Complex
  - SMO (Support Vector Machine)
  - Neural Networks

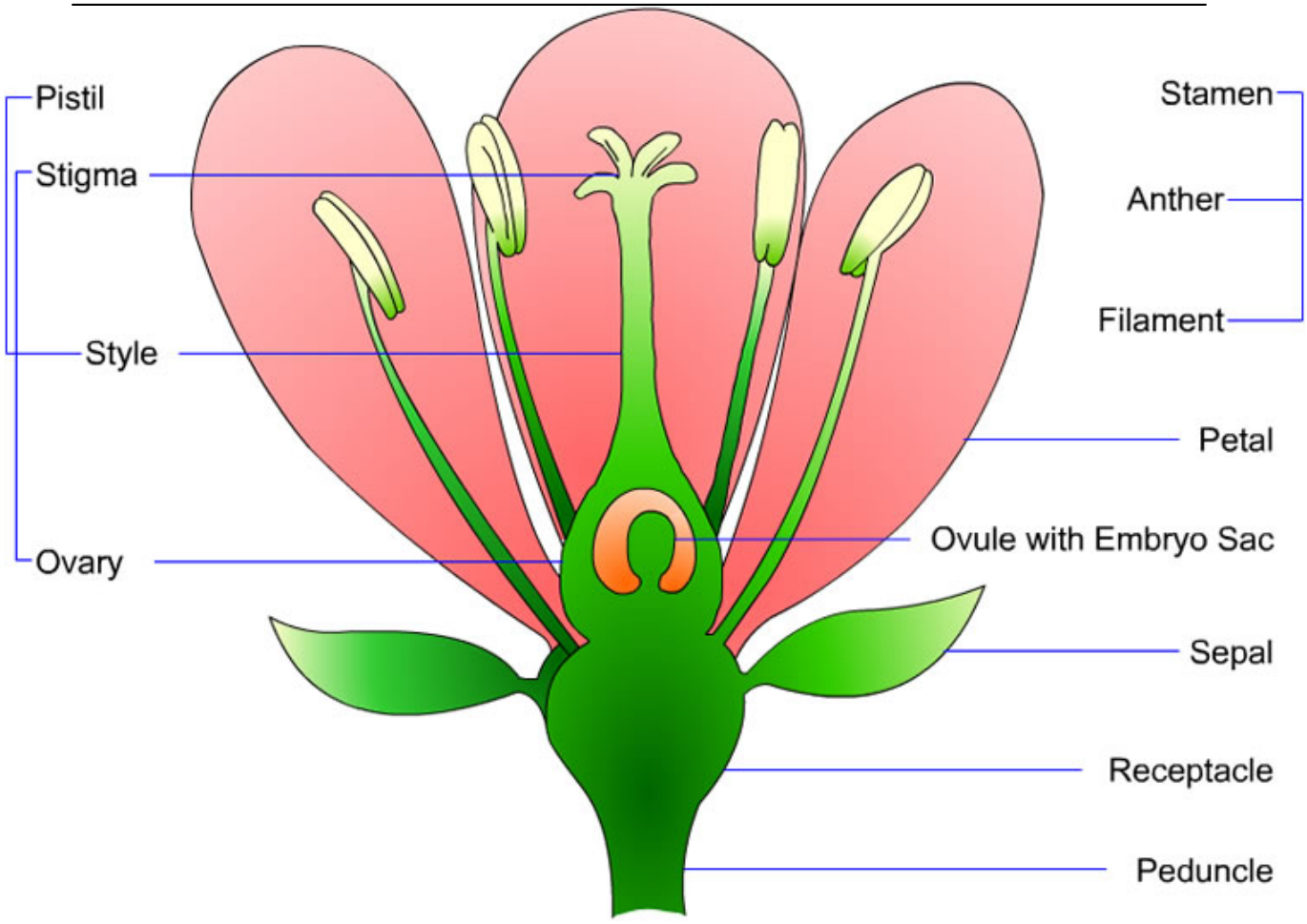
NO FREE LUNCH THEOREM

- There is no one best classification algorithm over all data sets

- Experience has shown that support vector algorithms are generally a bit more accurate on many data sets.

- Recently XGBOOST seems to have the edge.

IRIS FLOWER


Pistil
Stigma
Style
Ovary
Stamen
Anther
Filament
Petal
Ovule with Embryo Sac
Sepal
Receptacle
Peduncle

ARFF FILE OF IRIS DATA

[ARFF: Attribute-Relation File Format)]

```
@RELATION iris
% R.A. Fisher, 1936
@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@ATTRIBUTE class        {Iris-setosa,
                         Iris-versicolor,
                         Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
.
.
```

50 examples of each class, 150 in total.

NEAREST NEIGHBOUR CLASSIFIER

- Training algorithm
  - None
- Deployment stage (Testing)
  - Load the training data into memory
  - Load the unseen example
  - For each training example
    Compute distance to unseen example

$$||\vec{x} - \vec{y}|| = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

    m is the number of attributes
  - Return the class of the example with shortest distance

WEKA IB1

```
=== Run information ===
Scheme:weka.classifiers.lazy.IB1
Relation: iris
Instances:150
Attributes:5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

IB1 classifier

Time taken to build model: 0.02seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Total Number of Instances               51

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  0  0 |  a = Iris-setosa
  0 19  0 |  b = Iris-versicolor
  0  2 15 |  c = Iris-virginica
```