Remotely Sensing Cities and Environments

Lecture 6: Classification

02/02/2022 (updated: 14/02/2023)

a.maclachlan@ucl.ac.uk

andymaclachlan

andrewmaclachlan

Centre for Advanced Spatial Analysis, UCL

PDF presentation

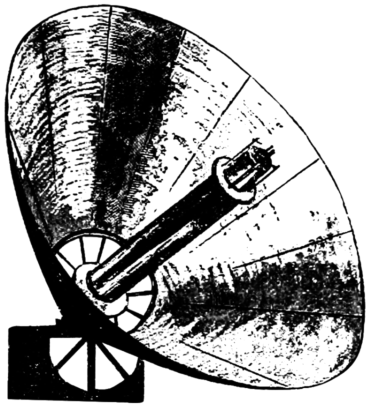
How to use the lectures

- Slides are made with xaringan
- 🔍 In the bottom left there is a search tool which will search all content of presentation
- Control + F will also search
- Press enter to move to the next result
- ✏ In the top right let's you draw on the slides, although these aren't saved.
- Pressing the letter o (for overview) will allow you to see an overview of the whole presentation and go to a slide
- Alternatively just typing the slide number e.g. 10 on the website will take you to that slide
- Pressing alt+F will fit the slide to the screen, this is useful if you have resized the window and have another open - side by side.

Lecture outline

Part 1: Review of how classified data is used

Part 2: How to classify remotely sensed data



Source:Original from the British Library. Digitally enhanced by rawpixel.

Let's look back at last week and see how some studies used classfied data

Urban expansion

Sensor

- Landsat

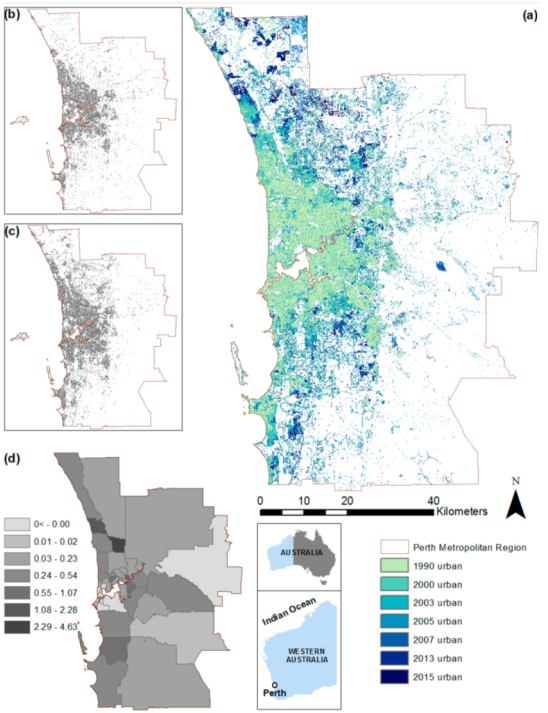


Figure 2. Urban expansion within the Perth Metropolitan Region (PMR) between 1990 and 2015. Vast urban growth has been observed in PMR with graduating colours exhibiting outward expansion (a); (b) and (c) exhibit static snapshots of urban extent from 2000 (b) and 2015 (c); whilst (d) depicts percentage of urban change per subnational administrative boundary (Local Government Area; LGA).Source:MacLachlan et al. 2017

Air pollution and LULC

Sensors

- Sentinel-3 Sea and Land Surface Temperature
- Sentinel-5 Precursor Major Air Pollutants

LULC transformation on air pollution, increase MAP (Major Air Pollutants) and LST

- Used regression...
- Honeycombing - hex grids for different sensor data

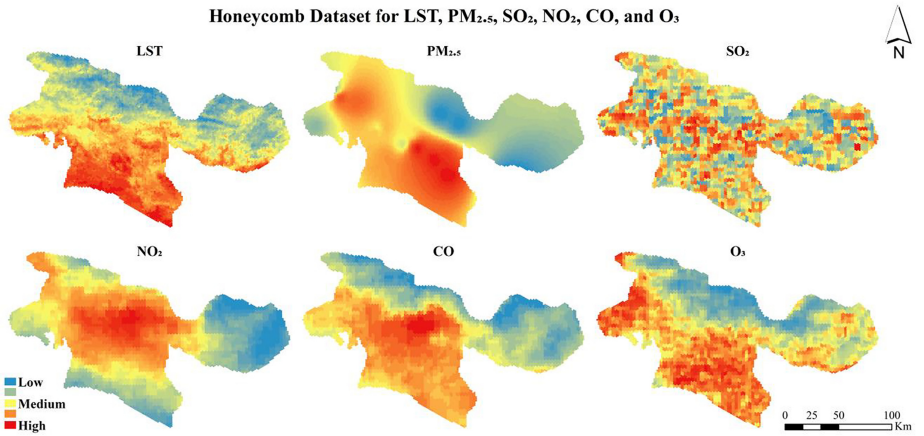


Fig. 2. The classified honeycomb dataset for LST, $PM_{2.5}$, SO_2, NO_2, CO, and O_3..Source:Fuldalu and Alta, 2021

Air pollution and LULC 2

> There is no doubt that the LULC distribution significantly affects the MAP and LST. Therefore, to determine this relationship the latest LULC distribution shape-file was acquired from the National Cartographic Center of Iran (NCC)
>
> To figure out the impact of LULC on the LST and MAP (Major Air Pollutants) the following statistical comparison perform [summary stats - mean, min etc)], the LULC was chosen as an independent variable whereas LST, $PM_{2.5}$, SO_2, NO_2, CO, and O_3 are considered as dependent variables

Although this wasn't used in regression...that was just for the scatter plots of the variables...

But we have classified data (or we might) from a national center

- although no data is given
- no accuracy or method provided

Air pollution and LULC 3

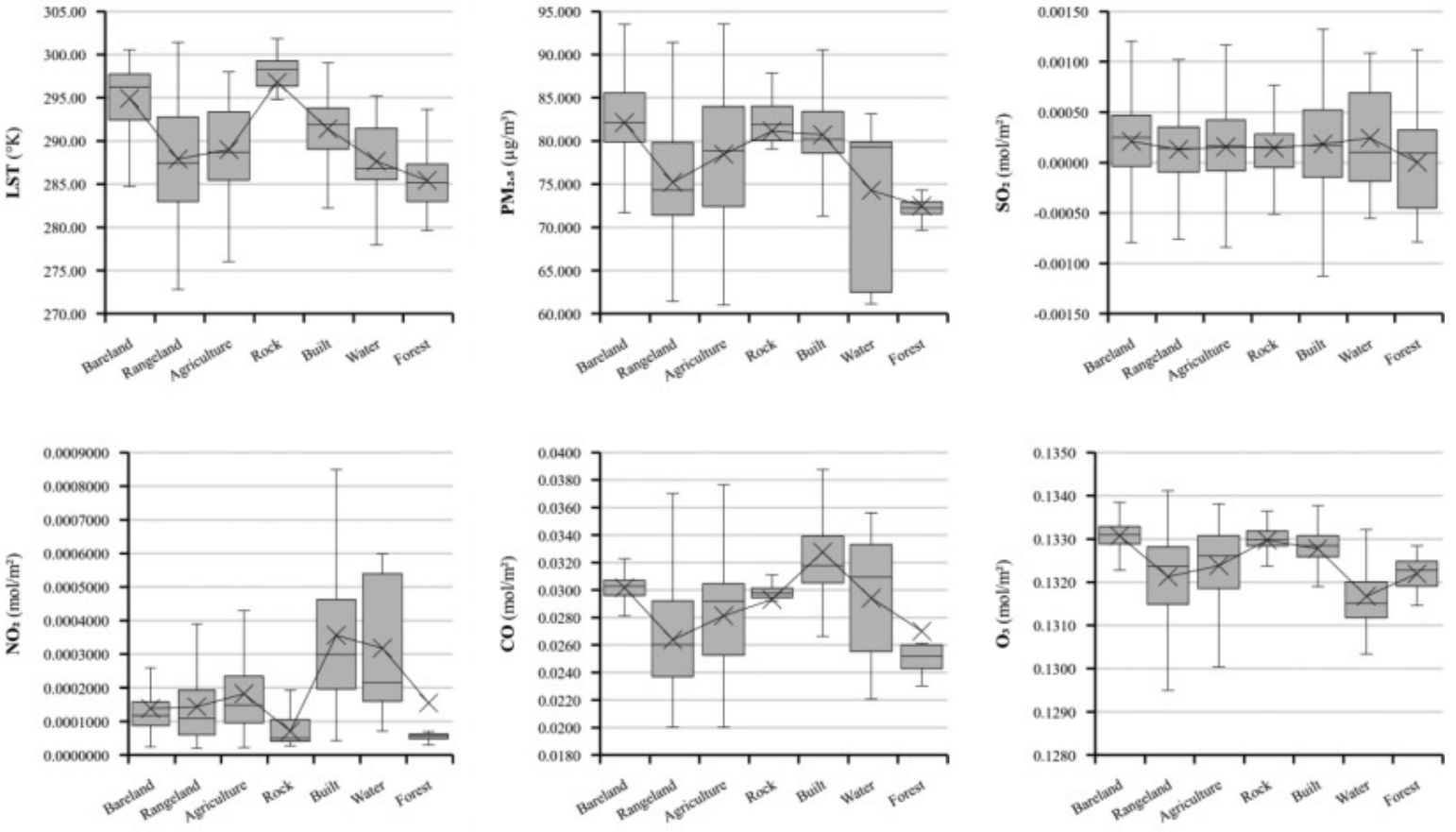


Fig. 3. Box-and-whisker plots between the LULC and the LST, $PM_{2.5}$, SO_2, NO_2, CO, O_3. Source:Fuldalu and Alta, 2021

Air pollution and LULC 4

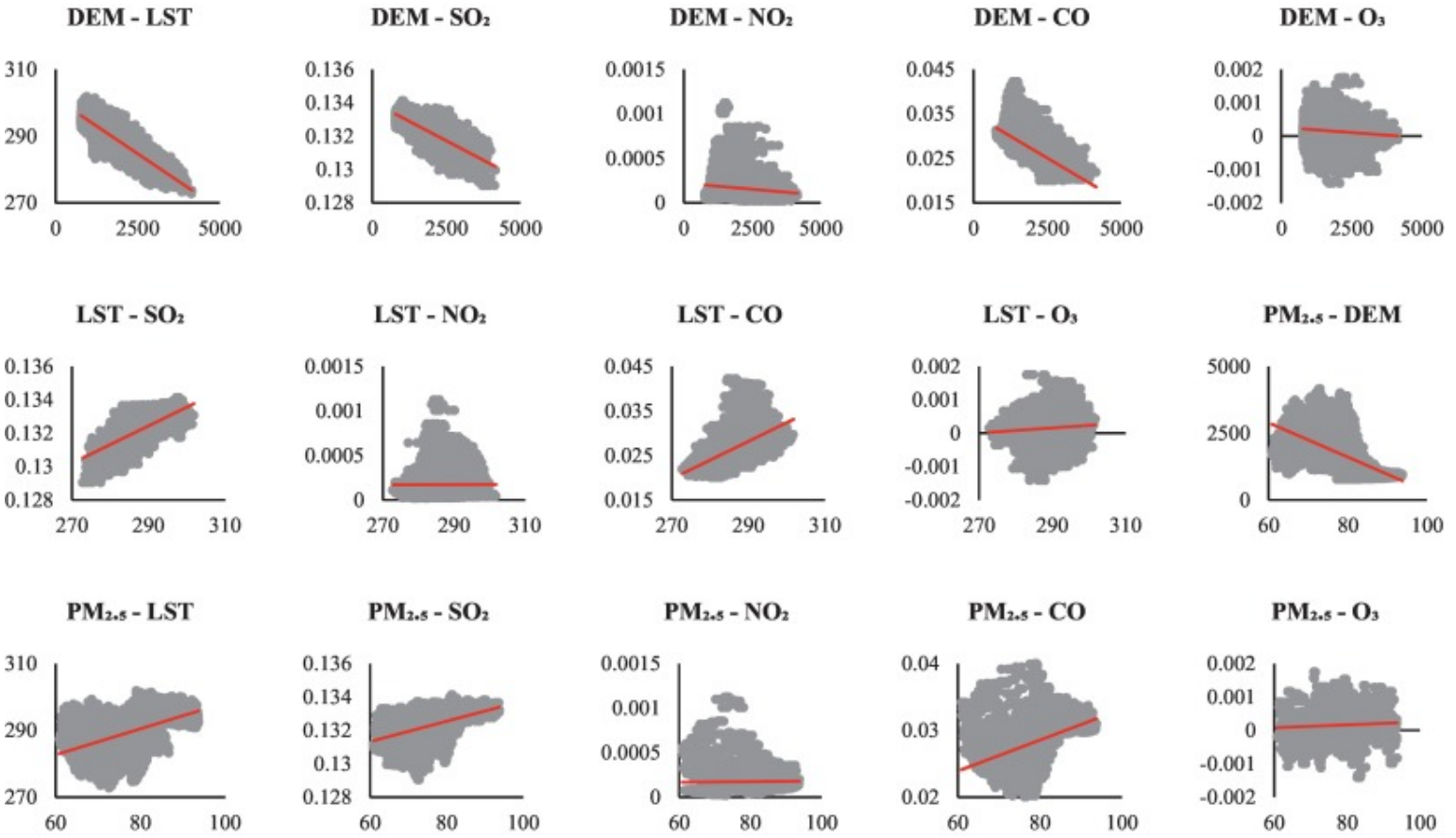


Fig. 4. The Scatter-plot among the DEM, LST, $PM_{2.5}$, SO_2, NO_2, CO, and O_3. Source:Fuldalu and Alta, 2021

Urban green spaces

> Our results show that the techniques are hybrid methods (37 cases), followed by object-based image analysis (29 cases), land cover indices (20 cases) and fraction methods (16 cases)

acronyms

Inventory and assessment (Inv_Ass);Biomass and carbon (BC);Change detection (CD); Ecosystem services (ES):Overall UGSs mapping (OUGS);Species mapping (Spe);Three-dimensional modeling (TDM).

Google refers to Google Earth products; High spatial resolution (Hig); High spatial resolution & Medium spatial resolution (Hig_Med); Hyperspectral (Hyp); LiDAR(Li); LiDAR & High spatialresolution (Li_Hig); LiDAR & Hyperspectral (Li_Hyp); Medium spatial resolution(Med).

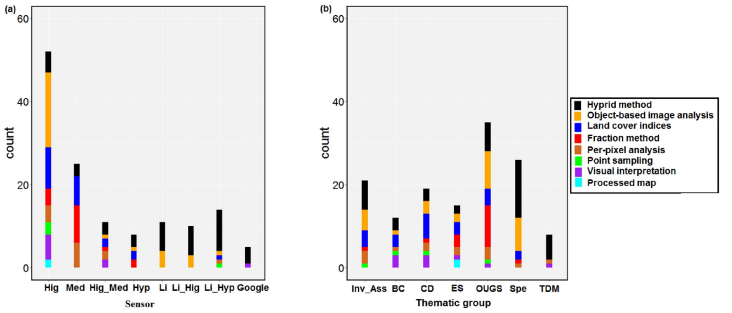


Fig. 4. Different techniques to characterize UGSs: (a) frequency of use of techniques according to type of remotely-sensed data, and (b) frequency of use of techniques according to application area. Source:Shahtahmassebi et al. 2021

Urban green spaces 2

Different sensors used for different mapping purposes, but can be mixed and matched (used intechangably)

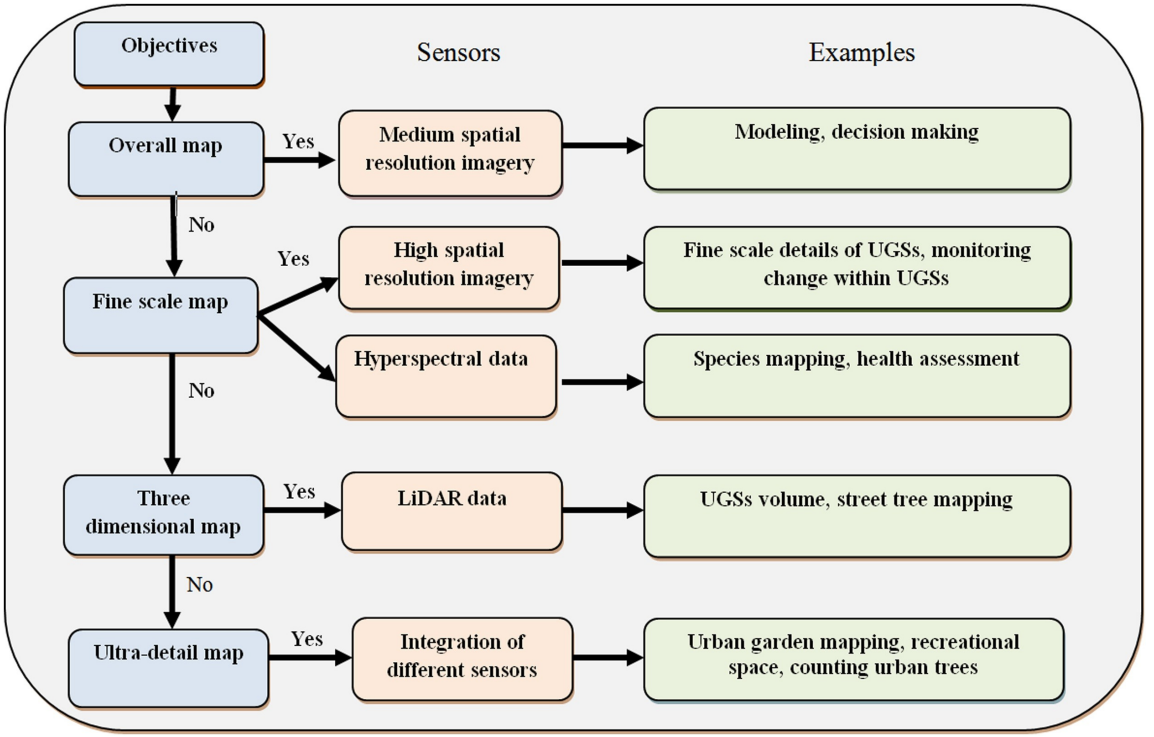


Fig. 5. Different techniques to characterize UGSs: (a) frequency of use of techniques according to type of remotely-sensed data, and (b) frequency of use of techniques according to application area. Source:Shahtahmassebi et al. 2021

Monitoring forests + illegal logging

- Leonardo Brito became chief of police at the Police Specialized in Crimes Against the Environment (DEMA) in Brazil's Amapá stated, he noticed that the department hardly ever investigated environmental crimes
- 2 employees, two vehicles, a boat and a drone, which collects only 20 minutes of footage at a time, to patrol an area of forest the size of Nepal.
- PRODES and DETER = annual data or 250m resolution
- Used Global Forest Watch produced by Hansen et al. in GEE

Environmental chief police Leonardo Brito and his team examine a deforested area. Image courtesy of DEMA-AP

Monitoring forests + illegal logging

- Uses the app version - Forest Watcher

> Brito said that since they starting using the app, Amapá's environmental police have been able to detect 5,000 areas of deforestation in the state, both legal and illegal. He adds that every day he sees new locations to add to the ever-growing list.

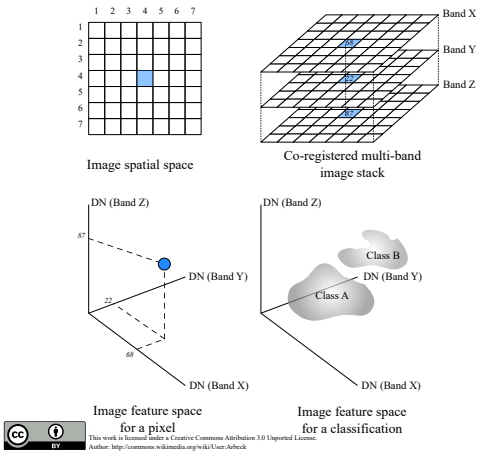
Trying to clear small patches to avoid detection!

Image courtesy of DEMA-AP.

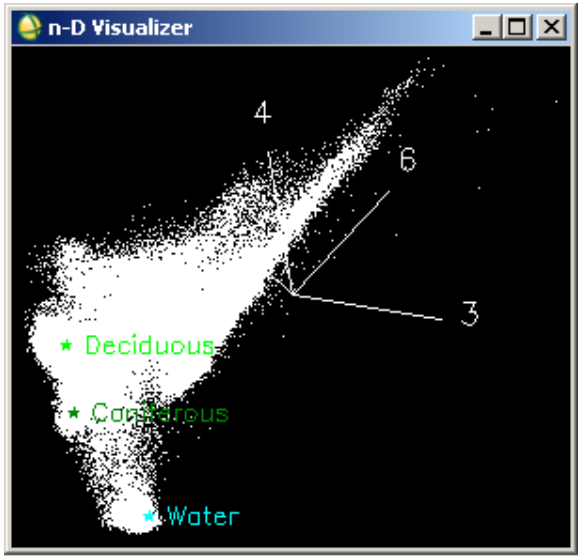
Monitoring forests + illegal logging

> "a more generalized feature space"

- Feature space = scattergram of two bands (or things that have been made into bands)
- Can be used for very basic classification - selecting the values that represent land cover



Feature space. Source:Wikimedia commons 2022

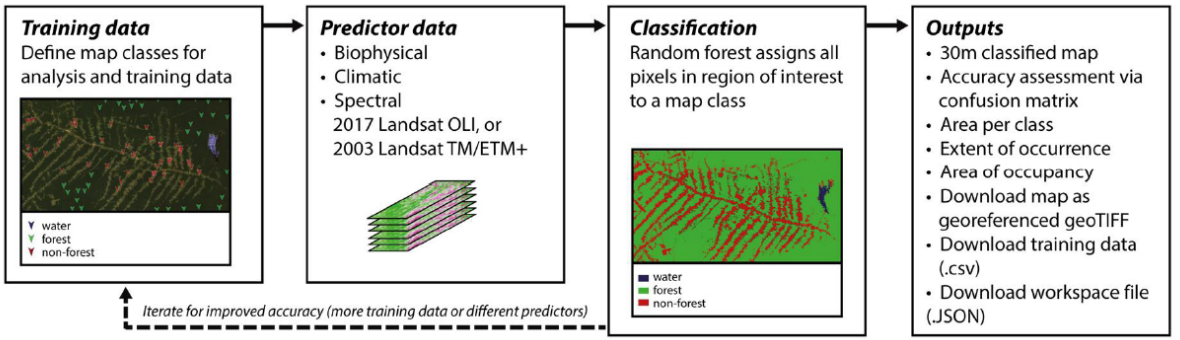


Spectral curves on the scatter plot. Source:50northspatial

Monitoring forests + illegal logging

- Training data (in supervised machine learning)

> Training data to relate to the Landsat metrics were derived from image interpretation methods, including mapping of crown/no crown categories using very high spatial resolution data such as Quickbird imagery, existing percent tree cover layers derived from Landsat data (29), and global MODIS percent tree cover (30), rescaled using the higher spatial resolution percent tree cover data sets



REMAP method. Source:UN-SPIDER

Monitoring forests + illegal logging

- Classification (supervised or unsupervised)

> Decision trees are hierarchical classifiers that predict class membership by recursively partitioning a data set into more homogeneous or less varying subsets, referred to as nodes

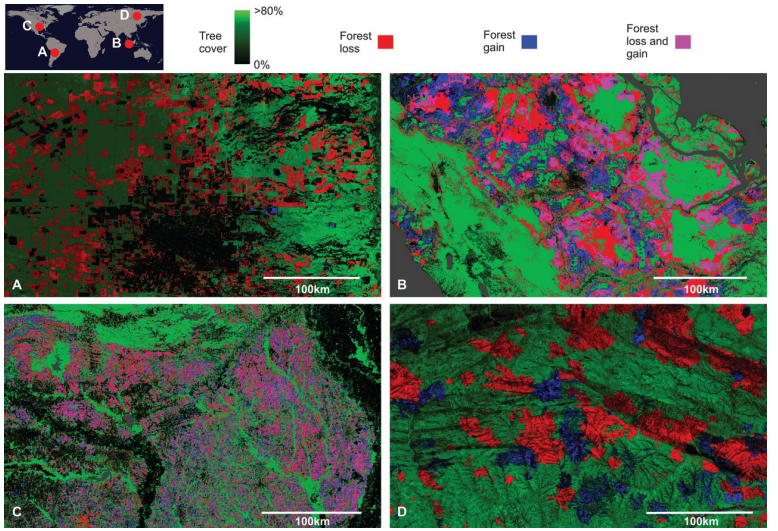


FIG. 2 Regional subsets of 2000 tree cover and 2000 to 2012 forest loss and gain.(A) Paraguay, centered at 21.9°S, 59.8°W; (B) Indonesia, centered at 0.4°S, 101.5°E; (C) the United States, centered at 33.8°N, 93.3°W; and (D) Russia, centered at 62.1°N, 123.4°E.
Source:Hansen et al. 2013

Used in Brazil to target illegal logging

Forest fires

- Dates back to the most cited paper on the topic
  - "Application of remote sensing and geographic information systems to forest fire hazard mapping", Chuvieco and Congalton 1989.

Used:

- **Sensor** Landsat TM 1984
- vegetation, elevation, slope, aspect and road/ house proxmity = fire hazard map compared to burned map from Landsat
- Did a weighted overlay of the layers - giving hazard value of 0 to 255, some layers had assigned values (e.g. aspect of 90-180 a value of 0)
- **Vegetation was from a classified Landsat TM image** - classified 16 categories
- No accuracy assessment
- I assume the manually delineated the burned area pixels

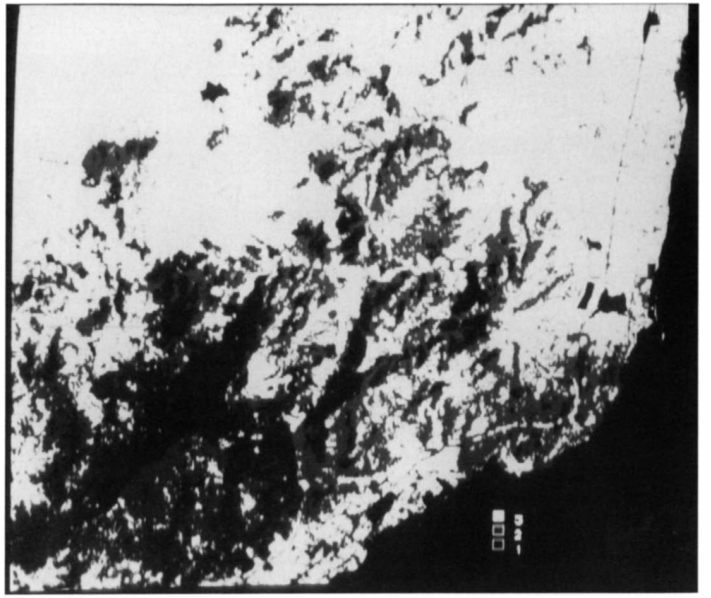


Figure 8. Hazard index map. For display purposes, the whole range of hazard index values has been divided in three categories: 1) high hazard (1–100); 2) medium hazard (101–200); 3) low hazard (201–255).

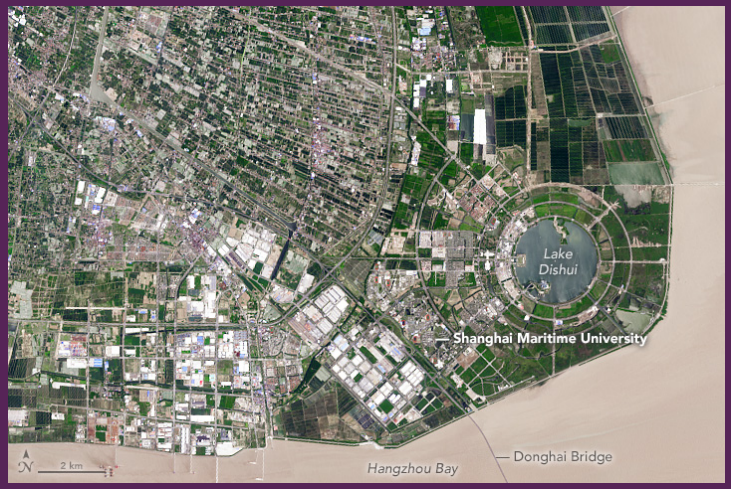
Source:Chuvieco and Congalton 1989

In some form all these studies extracted Land Cover from EO data

But how can we do that

How do you do that given some imagery ?

Source:NASA, acquired April 23, 1984

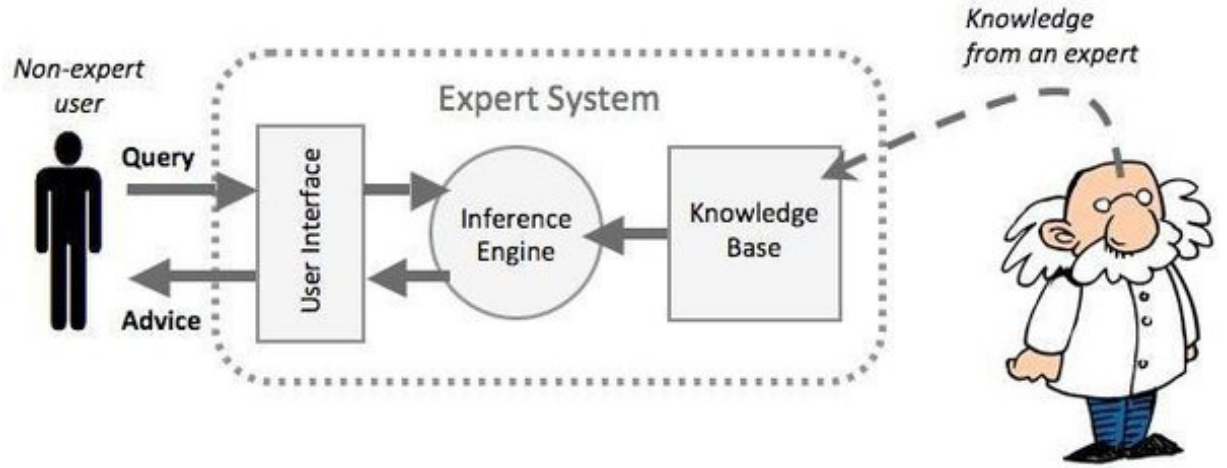


Source:NASA, acquired July 20, 2016

inductive learning = given context we can use experience to make judgement

Expert Systems

> a system that uses human knowledge to solve problems that normally require human intelligence



Source:Aftab Alam

- Knowledge Base = Rules of thumb, not always correct
- Inference Engine = Process of reaching a conclusion and the expert system is implemented

This is different to an algorithmic approach = code to solve a solution and when the problem changes so does the code. See Jensen p.433

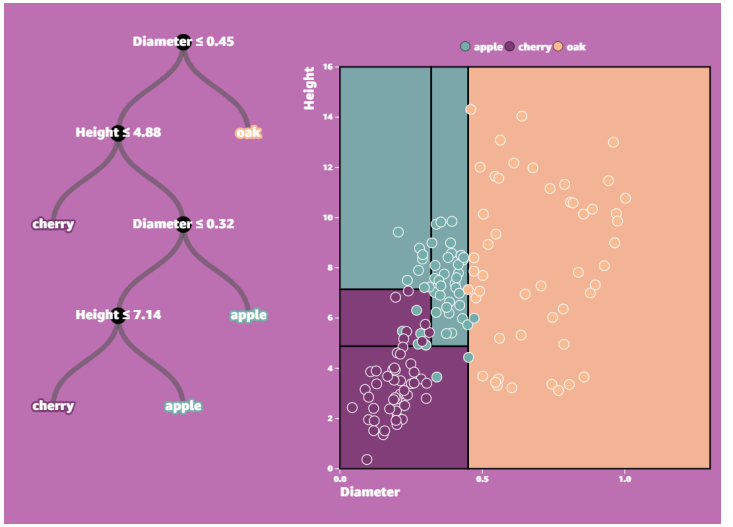
Expert Systems 2

The problem is how can a computer replicate human knowledge...

Q: Tell a computer how you arrived at the decision to wear the clothes you have on today or what you had for lunch yesterday ?

You might try and represent your knowledge through a series of decisions = **knowledge representation through a decision tree**

If you collected data on this you might be able to draw some conclusions...



From the diameter and height of a tree trunk, we must determine if it's an Apple, Cherry, or Oak tree. Source:Machine Learning University explain

Machine learning = science of computer modeling of learning process

When humans have some generalizations we can reach a logical assumption or conclusion = inductive learning.

Machine learning this is a search through all the data to explain the input data and can be used on new input data

What city am i in?

Population of 5.3 million

Median house price $1,116,219

Hemisphere: Southern

Continent: Australia

Landmark: Opera house

Is linear regression machine learning?

Yes, the model finds the best fit between independent and dependent variables

You are fitting a model to some data which could be used for prediction...

Classification and regression trees (CART)

Comprised of

classification trees

- classify data into two or more discrete (can only have certain values) categories
- For example, should you play golf today?
    - temperature
    - rainfall
    - wind
    - saturation

regression trees

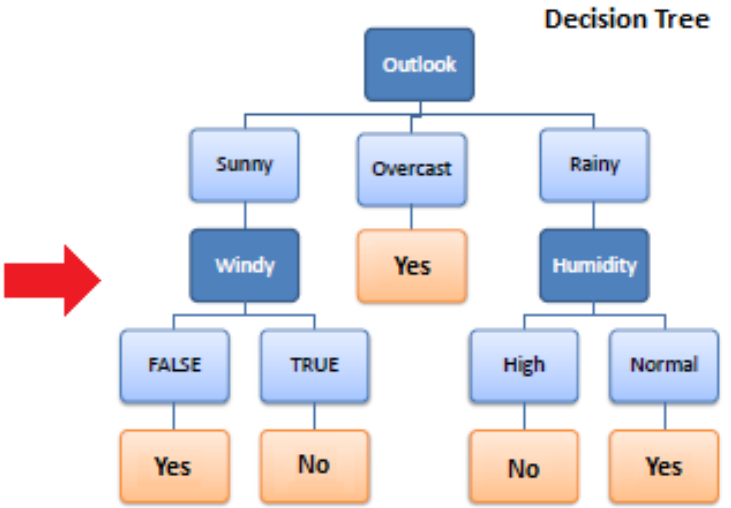
- predict continuous dependent variable
    - GCSE scores! the timeless example
    - Linear regression does work as...not a linear relationship
    - Large residuals

Classification and regression trees (CART)

- Classification tree

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

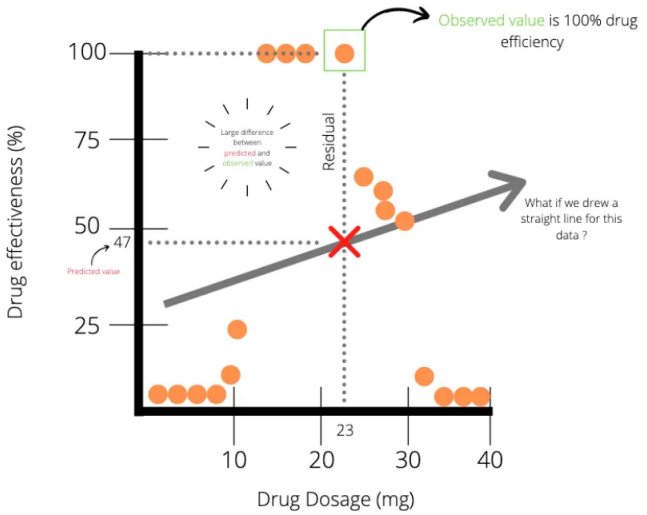


Decision Tree - Classification. Source:An Introduction to Data Science, Dr Saed Sayad

Classification and regression trees (CART)

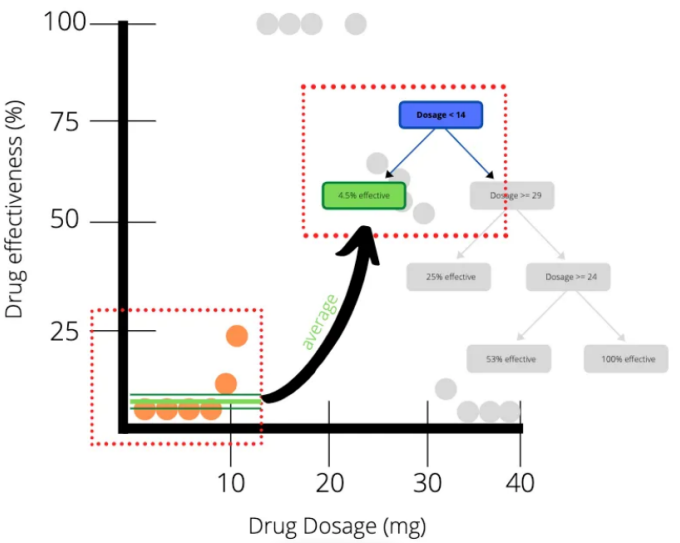
- Regression tree - subset the data into smaller chunks

Linear regression doesn't fit



How do Regression Trees Work?. Source:Luka Beverin

So...subset the data

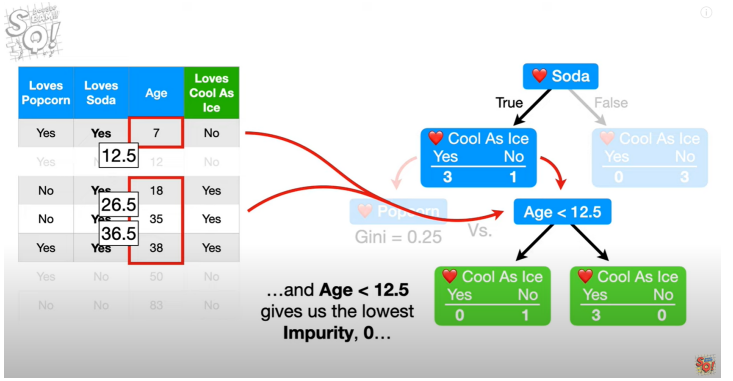


How do Regression Trees Work?. Source:Luka Beverin

Classification and regression trees (CART)

- When we create a decision tree the final leaves might be a mixture of the categories = **impure**

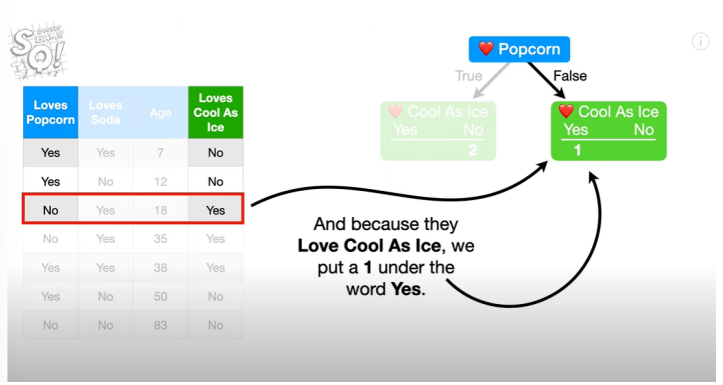
- We quantify this with the Gini Impurity (worked example on next slide):
  - 1-(probability of yes)^2-(the probability of no)^2
  - weighted based on numbers
- The one with the lowest impurity goes at the top of the tree to start the decision making...**the root**
- We then use the Gini impurity at each **branch** to split the nodes further
- Once we don't need to split these turn into **leaves** and the output has **the most votes**



Source:StatQuest

Gini impurity in more detail...

- How do we decide what data to start the tree with ?
- Does someone **who loves popcorn or soda** love the song cool as ice?
- If we have a yes and no we phrase this an impure leaf...and we quantify this with the Gini Impurity



Source:StatQuest

- Gini impurity= 1-(probability of yes)^2-(the probability of no)^2

$$1 - (\frac{1}{1+3})^2 - (\frac{3}{1+3})^2$$

Gini impurity for popcorn = true is 0.375 vs false is 0.444

But they have different numbers of people, so we take weighted average for the variable...

weight for left = people in leaf (4) / total in both leaves (7) *impurity (0.375) weight for right = people in leaf (3) / total in both leaves (7)* impurity (0.444)

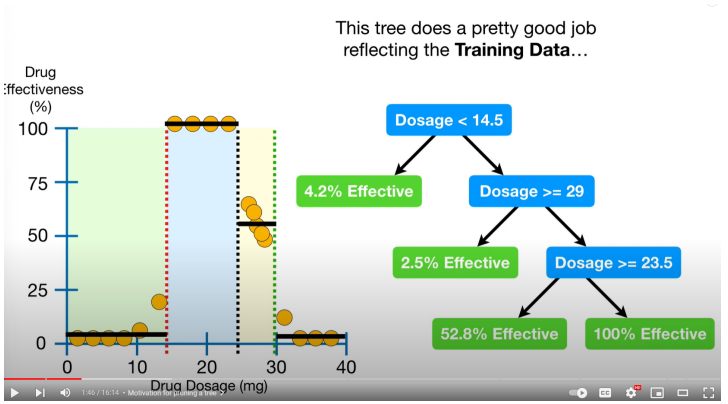
Add together, so impurity for loves popcorn is 0.405

lowest impurity wins

Someone new comes along ...run them (or the data) through the tree

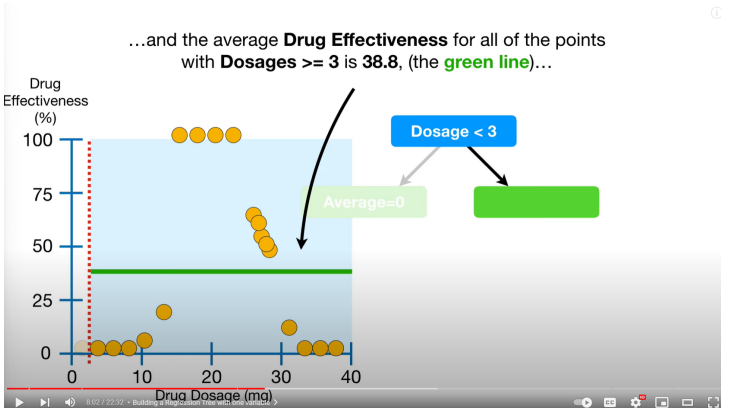
Regression trees

- Take and predict continuous values (e.g. amount of pollution)
- Classification trees take and predict discrete values (e.g. landcover)



Source:StatQuest

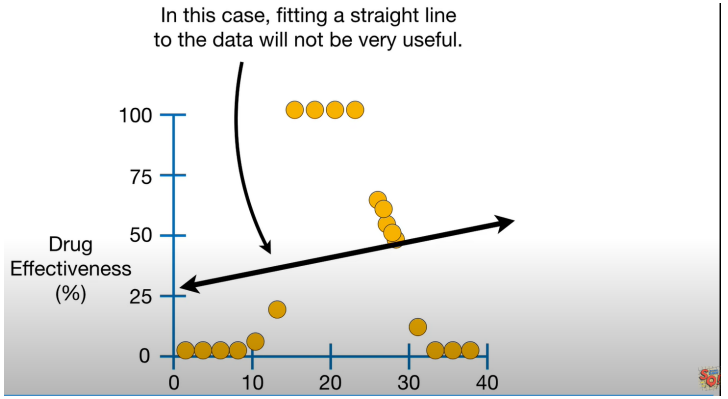
- How do we decide where to make the breaks in the data...**residuals** (like linear regression) for each threshold (which is a value on the x axis)



Source:StatQuest

Regression trees

What if linear regression doesn't fit the data? ...but we still wanted a numeric value



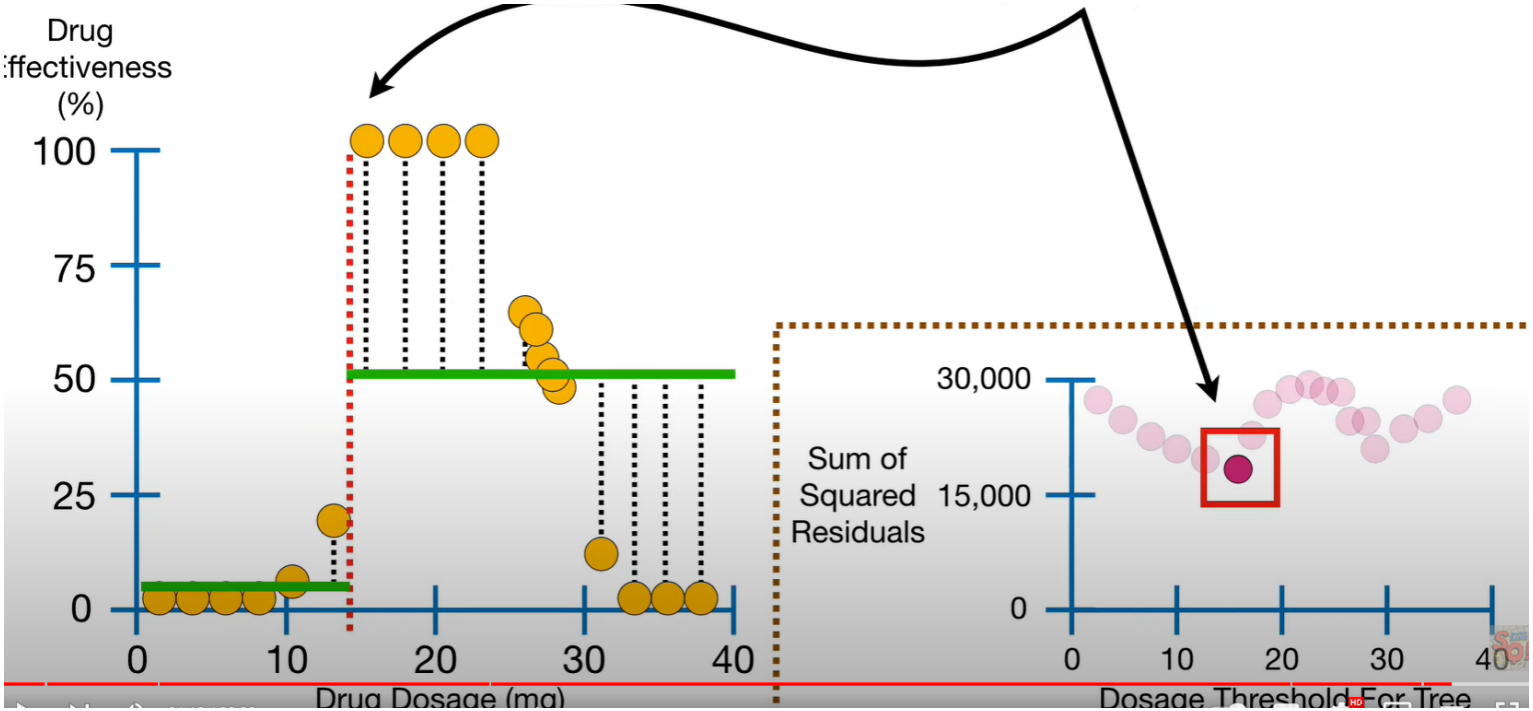
Source:StatQuest

We divide this up into sections based on thresholds (nodes) and calculate the sum of the squared residuals...

We can then check the SSR for different thresholds...**the one with the lowest SSR is the root of the tree to start with**...then repeat

To prevent over fitting we can set **a minimum number of observations before splitting the data again**.

Regression trees 2



Source:StatQuest

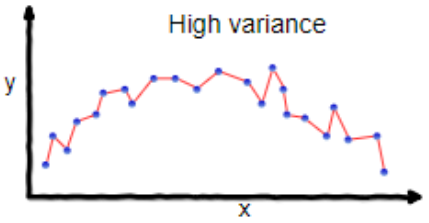
Regression trees 3

- We can do this with many predictor variables...we try different thresholds and calculate the sum of squared residuals (SSR) - e.g. age or gender
- The best sum of squared residuals (SSR) value **across all variables becomes the root**.
- Each predictor is then used in the process based on lowest sum of squared residuals (SSR)
- Each leaf **is a numeric value** not category like in classification trees.

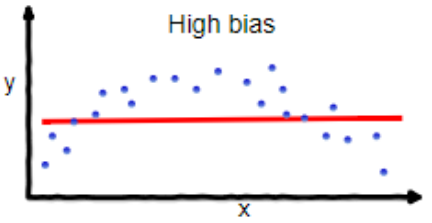
Overfitting

What if we have a leaf with just one person or one pixel value? = **overfitting**

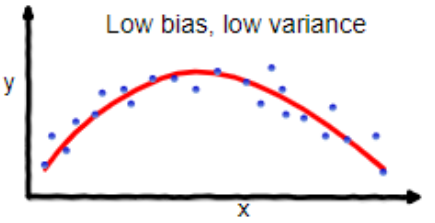
- Bias = difference between predicted value and true value = oversimplifies model
- Variance = variability of model for a given point = does not genearlise well



overfitting



underfitting



Good balance

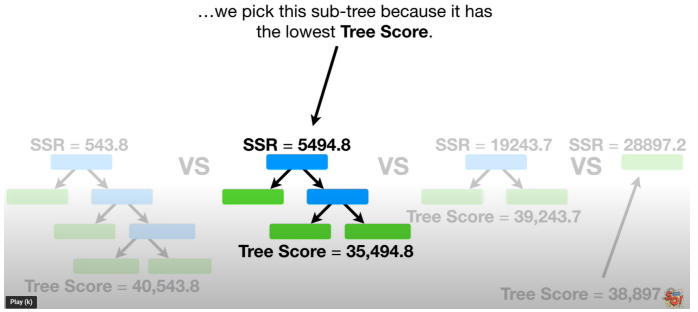
Source:Seema Singh

Overfitting 2

We can either:

- limit how trees grow (e.g. a minimum number of pixels in a leaf, 20 is often used)
- Weakest link pruning:
  - calculate the sum of the squared residuals (SSR, like linear regression) for each leaf (final decision)
  - use one less leaf, remove a leaf = **sub-tree**, SSR will get larger = **PRUNE**
  - Sum for the tree
  - Tree score = SSR + tree penalty (alpha) * T (number of leaves)

Overfitting 3



Source:StatQuest

- Alpha: use a full size regression tree (with **all** data), start with a value of 0 then increase **until pruning** (removing a leaf) gives lower **tree score** .
  - Divide the data into training and testing data using the alpha values from before
  - Calculate the SSR using the testing data...which alpha has the smallest SSR
  - repeat with different training and testing data (10 times)
- On average the value of alpha that gives lowest SSR from testing data is the final value.

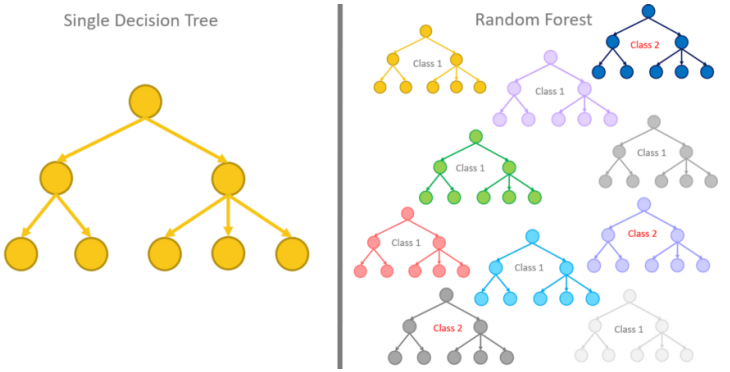
Decision trees aren't great with new data...

Many are better than one

Random Forests

- Grow many classification decision trees - **Many better than one**
  - Take our data and take bootstrap samples (same data can be picked many times)
  - Make decision tree from random number of variables (never all of them)
  - Next at the node take a random subset of variables again = **RANDOM**
  - Repeat!

- We get many, many different trees = a **forest**
- Run the data we have down the trees
- Which option gets more votes based on all the trees

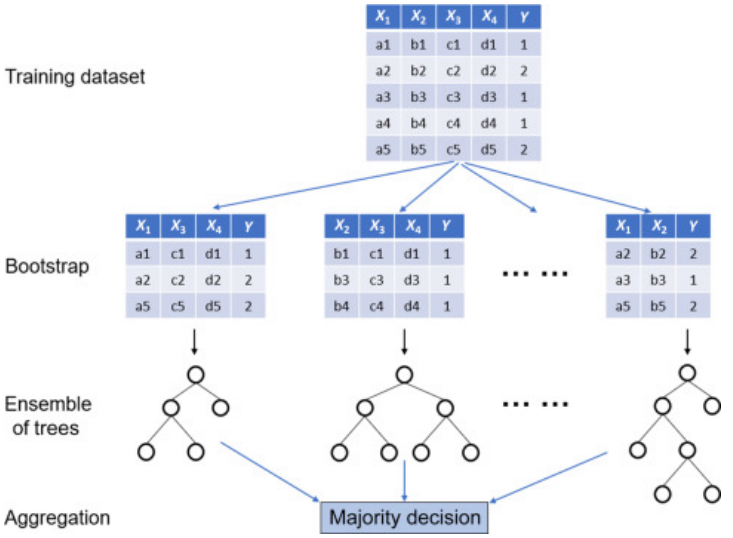


Source:Rosaria Silipo

Random Forests 2

- Bootstrapping (re-sampling by replacement data to make a decision = **bagging**
  - For each tree about 70% of the training data is used in the bootstrap, 30% is **left out of the bag (OOB)**
  - Test the OOB data in the forest where all the trees didn't use it
  - Most votes wins!
  - Repeat for all OOB samples
- proportion of OOB incorrectly classified = OOB error

- Often the number of variables per tree is calculated from square root of variables in the original data.



Random Forest and overview. Source:Science Direct

Random Forests 3

- No pruning - trees can do down to largest extent
- **Out of Bag Error**
  - all trees that didn't have the values (e.g. rows in the data) in
  - average prediction error - number of correct predicted/total
- Validation data
  - different from OOB
  - never included within the decision trees

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

Bootstrap sample

Then the last row that is "left out" in the original data (see the red box in the image below) is known as Out of Bag sample. This row will not be used as the training data for DT 1. Please note that in reality there will be several such rows which are left out as Out of Bag, here for simplicity only one is shown.

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

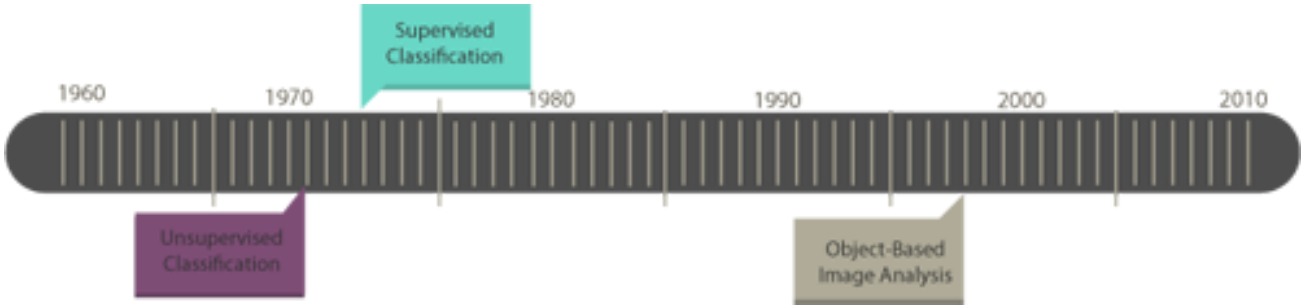
Out of Bag sample

Source:Navnina Bhatia

How do we apply this to imagery

Source:NASA, acquired July 20, 2016

Classification trends



Source:GISGeography

Image classification

- Turn every pixel in the image into one of a pre-defined categorical classification
- Either supervised or unsupervised classification procedure:

supervised

- Pattern recognition or machine learning
- Classifier learns patterns in the data
- Uses that to place labels onto new data
- Pattern vector is used to classify the image

Usually pixels treated in isolation but as we have seen - contextual (neighboring pixels), objects (polygons), texture

unsupervised

- Identify of land cover classes aren't know a priori (before)
- Tell them computer to cluster based on info it has (e.g. bands)
- Label the clusters

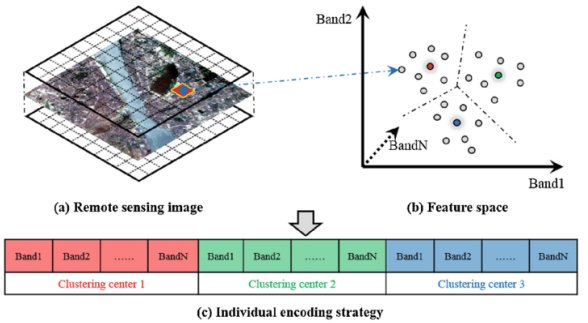
There are *generic* machine learning algorithms and remote sensing specific ones*

Unsupervised

Usually referred to as clustering / also k-means:

- Place points randomly or uniformly across spectral feature space or across the first PCA
  - Set the radius in spectral feature space at which new cluster to new started
  - Spectral distance to merge (within they are the same)
  - Number of pixels to be considered before merging
  - Max number of clusters
  - Clusters migrate over time - see Jensen page 404.

Repeat until N iterations or no allocations of pixels left.



Source:Yuting Wan

Unsupervised 2

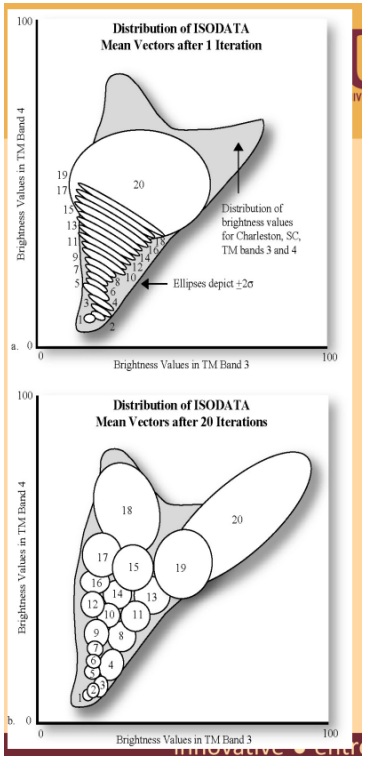
ISODATA

Same as k-means but adds:

- Any clusters have so few pixels = meaningless
- Clusters are so close they can be merged
- Clusters can be split - elongated clusters in feature space

Typically inputs can include:

- max clusters
- Max % pixels of class values that can be unchanged - stops
- Max times of iterations
- Min pixels per cluster
- Max standard deviation - then split the cluster
- Min distance between clusters (3)



Source: Jensen 2016 p.409 / Muhammad Zulkarnain Abdul Rahman

Unsupervised 3

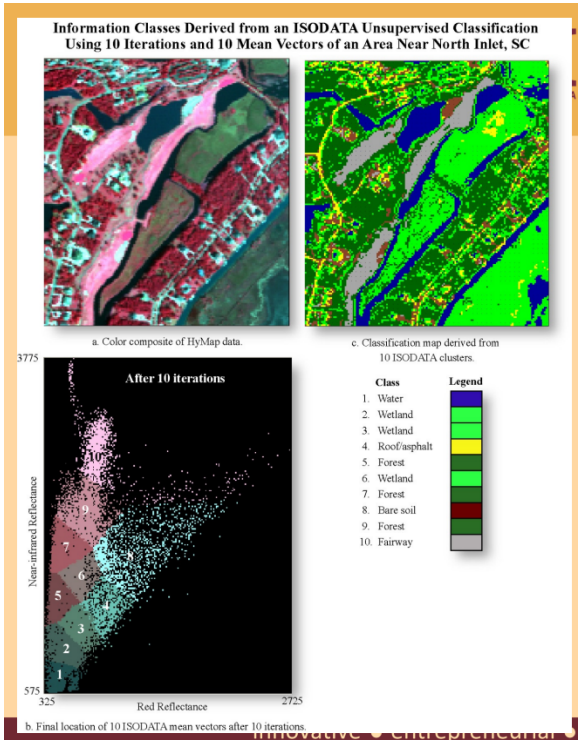
Cluster busting

ISODATA can create lots of clusters and it's difficult to assign meaning (e.g. landcover)

- Two types of landcover in the pixel
- distribution of mean vectors not good enough to differentiate them

Let's **bust those clusters!**

- Take the incorrect or difficult to label ones
- Mask them
- Perform a separate classification
- Repeat



Source:Jensen 2016 p.409 / Muhammad Zulkarnain Abdul Rahman

How does supervised differ from unsupervised...?

Supervised

Parametric (normal distribution) or non parametric (not normal)?

I would call most of these **"classical"** or **"traditional"** classifiers as they aren't used *much* now

Parametric

- Maximum likelihood

Non-parametric

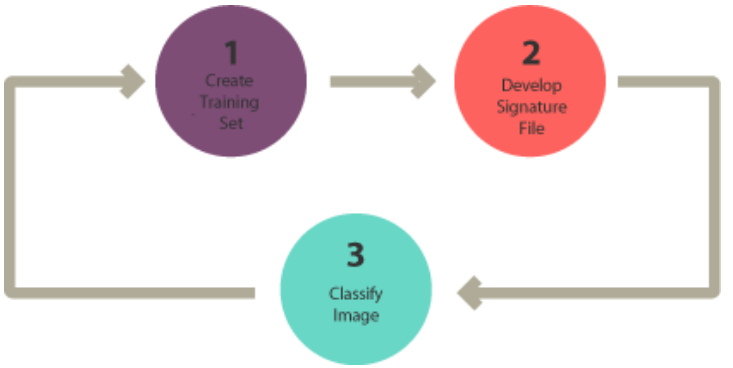
- Density slicing
- Parallelpiped
- Minimum distance to mean
- Nearest neighbor
- Neural networks
- Machine learning / expert systems*

- More recent work uses machine learning / expert systems(e.g. Support Vector Machine, Neural Networks) or spectral mixture analysis

Supervised 2

Same process for all:

- class definition
- pre-processing
- training
- pixel assignment
- accuracy assessment



Source:GIS Geography

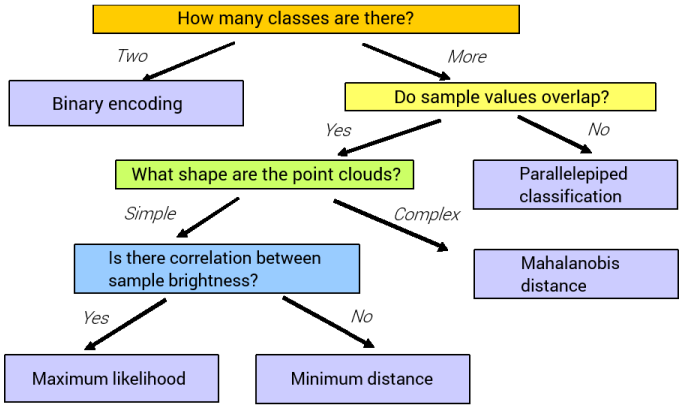
Here we will focus on two common methods - maximum likelihood and support vector machines

We have already covered decision trees and random forests

But...

Supervised 3

An approach to select a classifier...in most cases training samples will overlap...unless you select **spectrally pure endmembers** or use a **spectral library**.



Source:Pavel Ukrainski

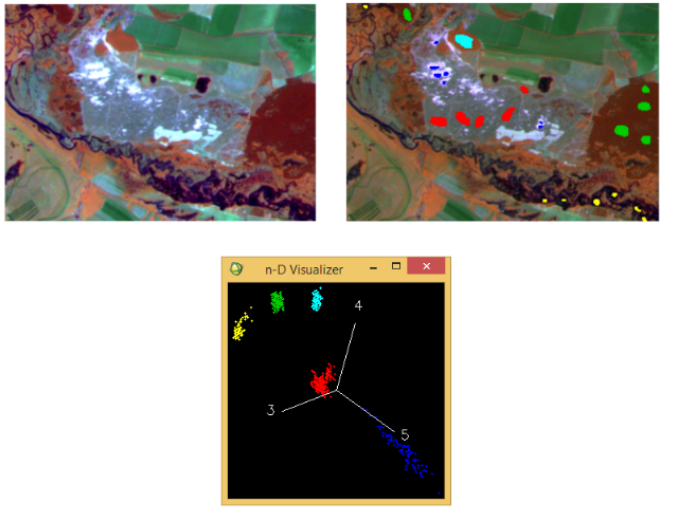


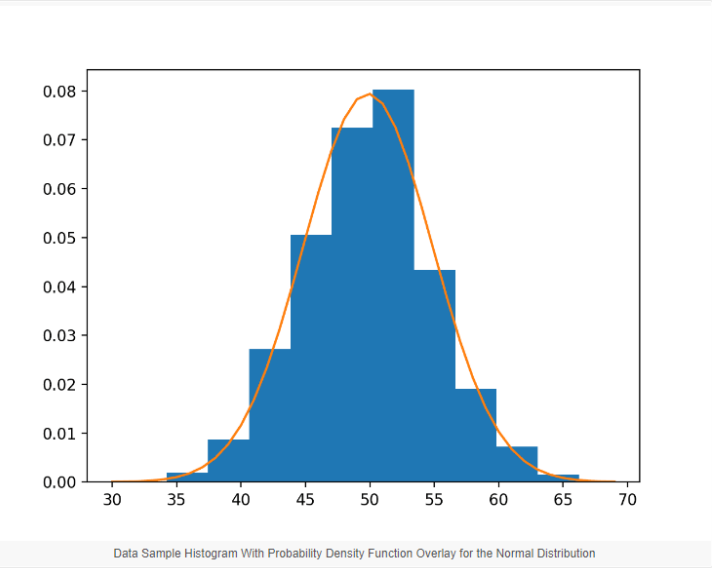
Figure 4. Example of when parallelepiped classification is most suitable

Source:Pavel Ukrainski

Maximum likelihood

Basics

- Decision rule classifier
- Uses probability
- Takes the image and assigns pixel to the most probable land cover type.
- You can set a probability threshold which means if a pixel is below it = no classification.

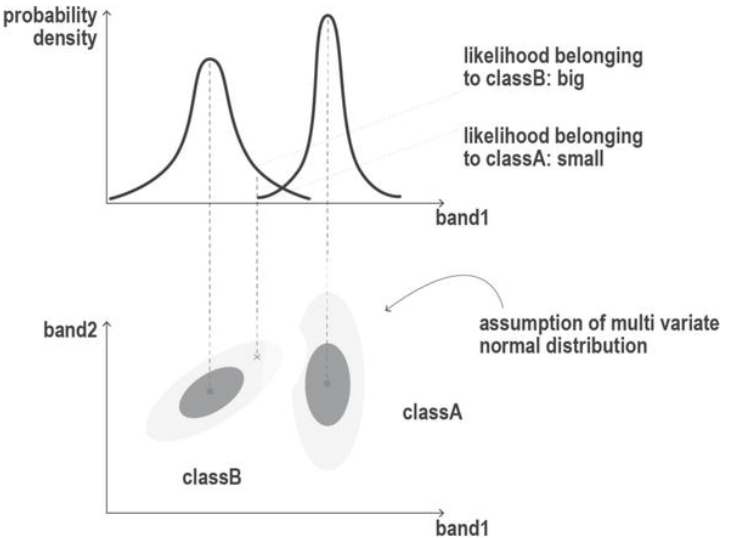


Source:Machine Learning Mastery

Maximum likelihood 2

Specifics

- From histogram to probability density function
  - mean and standard deviation of training data
- In imagery this is *n* dimensional multivariate normal density function - see Jensen p.399 for equation.
- Each pixel from image data is passed to the maximum likelihood rule > assigns landover to the largest product.



Source:Núñez et al. 2018 High-Resolution Satellite Imagery Classification for Urban Form Detection

Maximum Likelihood allows classification with **prior** pobablity information (e.g. 60% expected to be urban)

Usually we don't have this though

Terminology alert

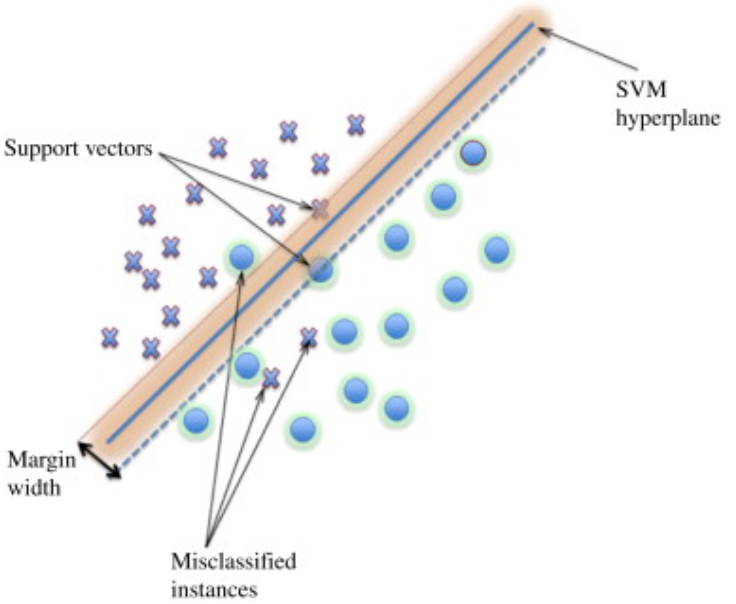
Pattern vector = all the band values per pixel (could include texture etc)

Also recall how we can fuse data from lecture 3

Support Vector Machine (SVM)

- Simply a linear binary classifier - like logistic regression!
- Maximum **margin** between two classes of training data
- Points on the boundaries are **support vectors**
- Middle margin is called the **separating hyperplane**

This can be thought of as training data for class a on one side and training data for class b on the other side, with band 1 on the x axis and band 2 on the y axis.



Source:Núñez et al. 2018 High-Resolution Satellite Imagery Classification for Urban Form Detection

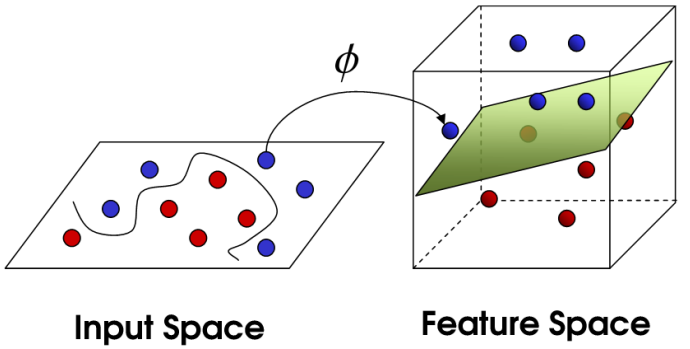
Support Vector Machine (SVM) 2

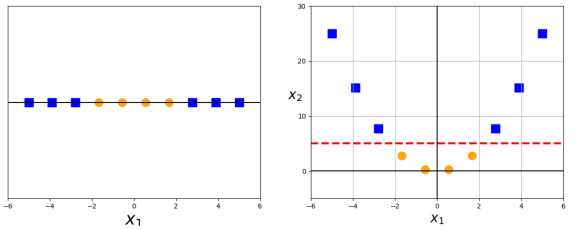
- Underlying theory is **structural risk minimisation**
  - Minimise error on unseen data with no assumptions on the distribution

Selectable:

- Type of kernel
- **C** controls training data and decision boundary maximisation plus margin errors. The bigger = narrower margin.
- **Gamma (or Sigma)** low = big radius for classified points, high = low radius for classified points
  - More on this next week...

- If they aren't linearly separable we can transform the data with the **kernel trick**
  - Apply some function to make them linearly separable





Source:Drew Wilimitis

Support Vector Machine (SVM) 3

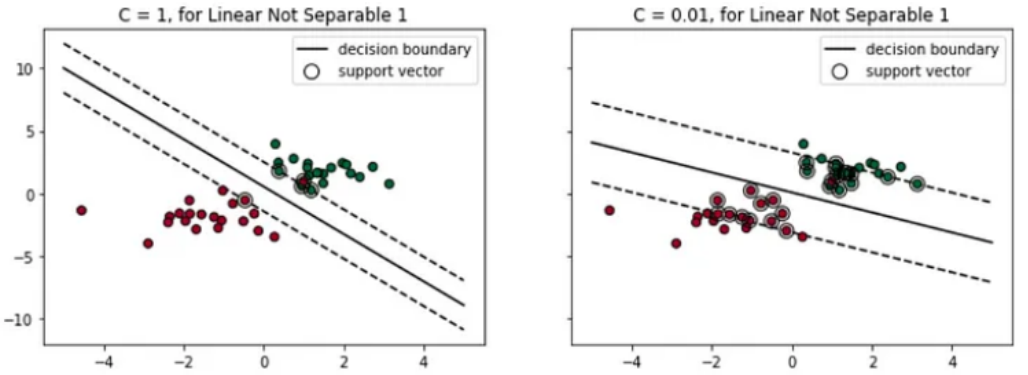
Hyperparameters like **C** and **Gamma (or Sigma)** control SVM wiggle...

In SVM we want to make sure each data set is on the correct side of a hyper plane

It does so through:

- Maximising the margin (the smallest residual)
- Minimising classified points..."**soft margin**"

Changing **C** changes the "slope" - consider more points

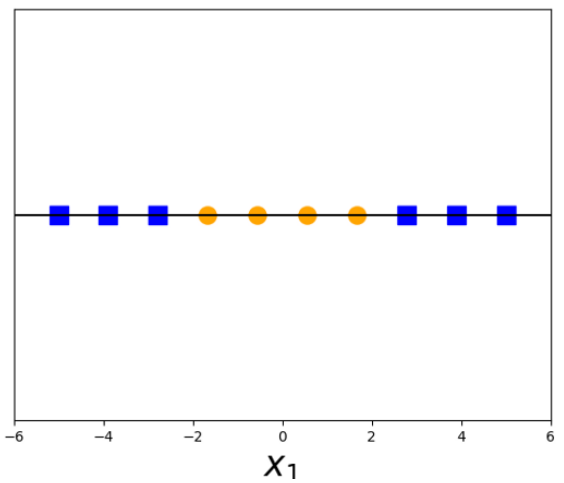


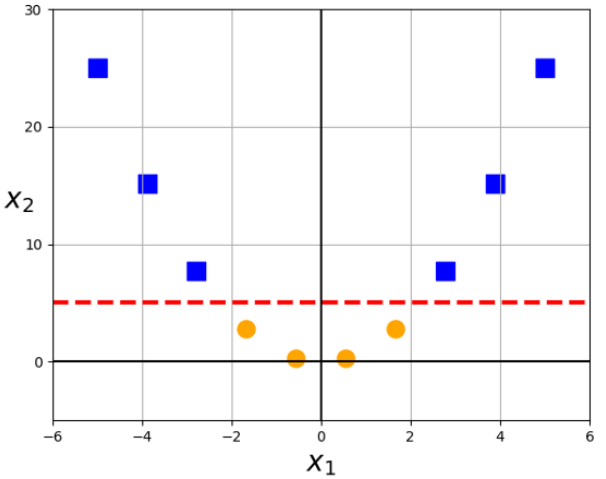
Source:Lilly Chen

Support Vector Machine (SVM) 4

As noted, if we don't have linearly separable we can transform the data with the kernel trick

....remember logging our data in linear regression?....similar idea...what is easier to separate...





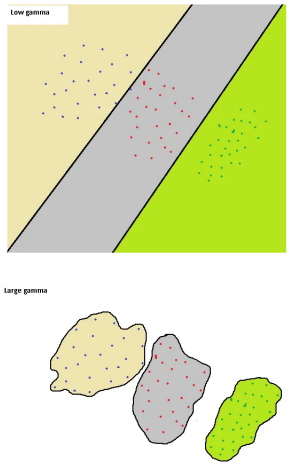
Source:Drew Wilimitis

Support Vector Machine (SVM) 5

only once we have transformed our data can we apply gamma

Gamma = controls the distance of influence of a training point...think back to CASA0005...min points in DBSCAN or weight matrix in spatial lag/error...

- Low value = lots of similarity
- High value = points near each other..

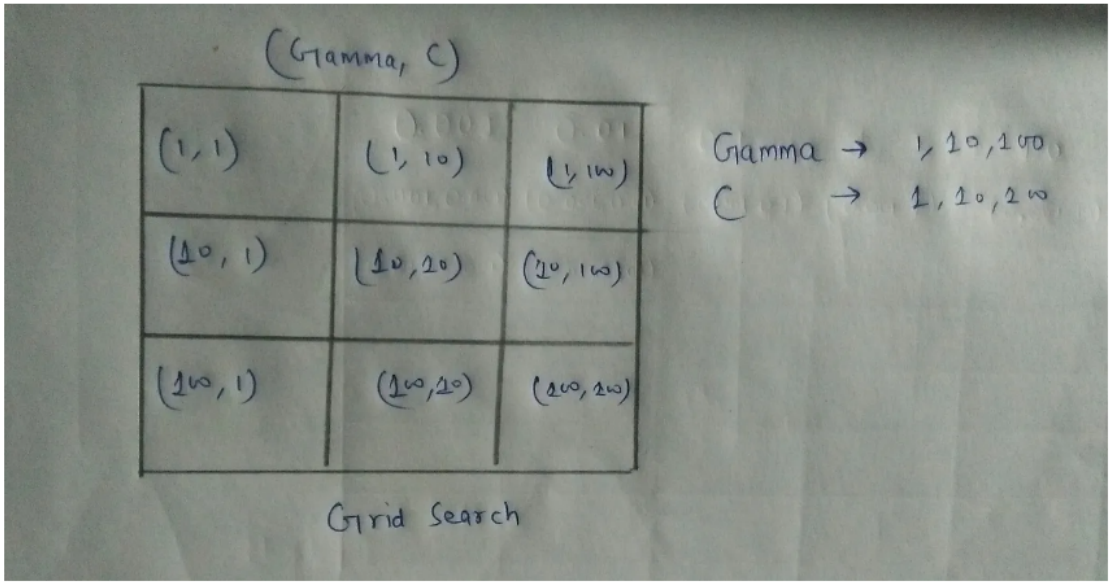


Source:Soner Yildirim

Support Vector Machine (SVM) 6

Q: how do we select the best values of **C** and **gamma**

A: We test them **all** (or all the ones you want to) using grid search and compare them to our testing data...the ones that give the best accuracy are selected...

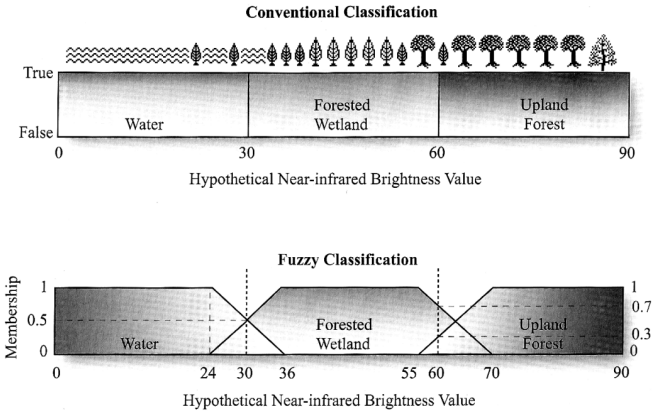


Source:A Man Kumar

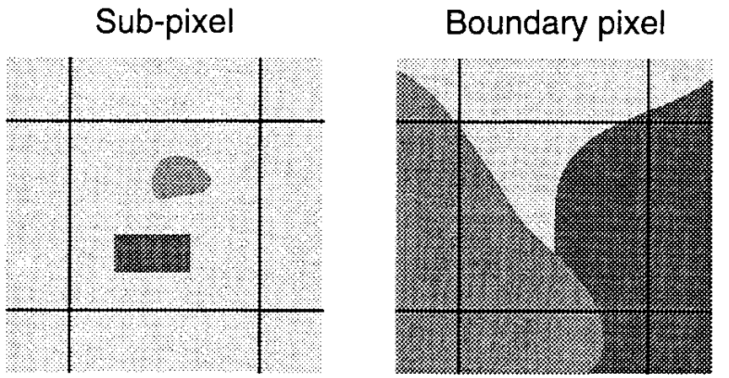
Some considerations...

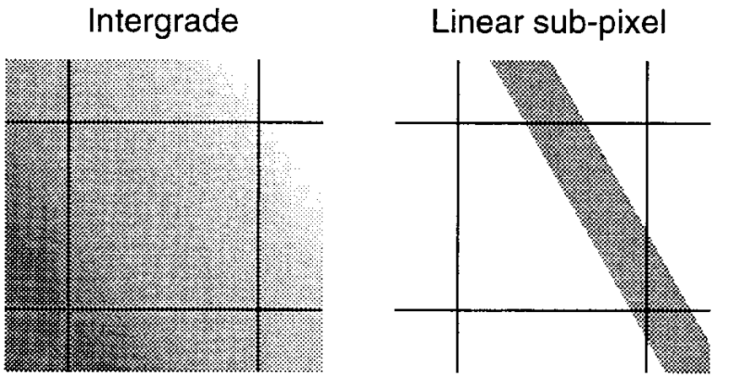
Do not just present a study that classifies imagery ...look back at the examples and how else LULC be used

Hard or soft?



Source:Jensen, page 413/Slides





Source:Peter Fisher

A. P. Cracknell, 2010, Review article Synergy in remote sensing-what's in a pixel?

Pixels or objects?

Source:Jakub Nowosad, supercells

Source:Jakub Nowosad, supercells

Average colour per segment. Source:Jakub Nowosad, supercells

Pixels or objects?

This stems from simple linear iterative clustering (SLIC)...which uses k-means clustering to create the superpixels. See Achanta et al. (2012)

Blackbox?

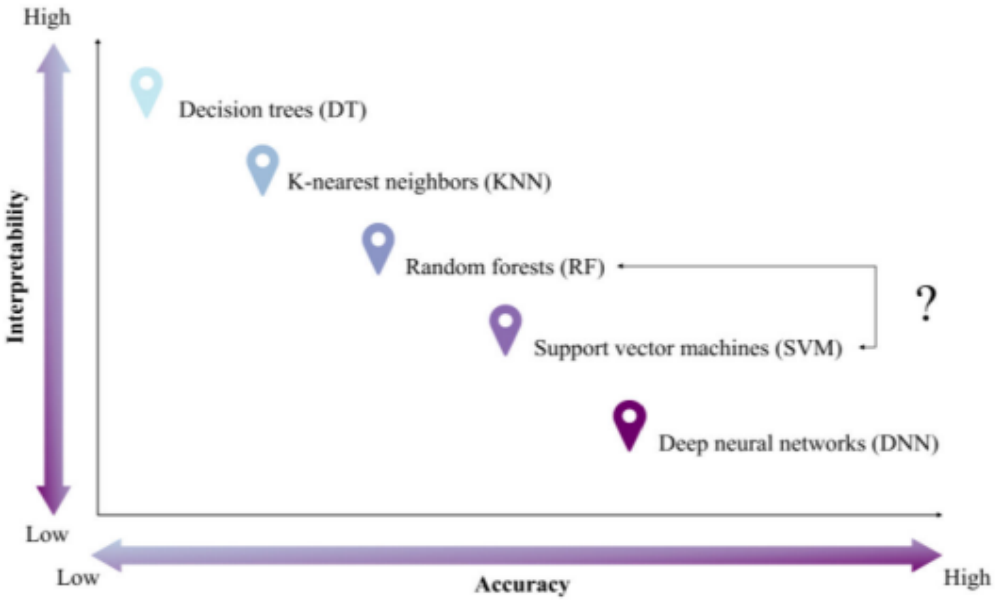


Fig. 1. Interpretability-accuracy tradeoff in machine learning classification algorithms.

Source:SHEYKHMOUSA et al. 2020 Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review

Summary