Finding new Correlations,
Surviving the Titanic

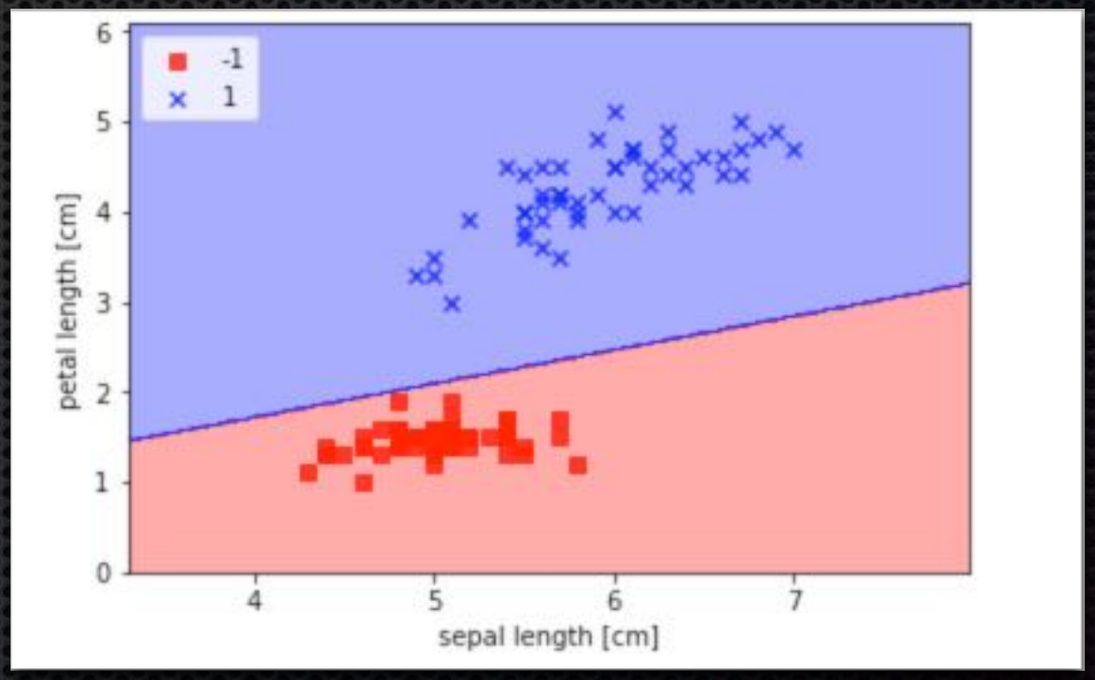
Sven-Patrik Hallsjö
University of Glasgow

University of Glasgow | Experimental Particle Physics

- Session 1
- Introduction to ML
- Using Sklearn
- Iris-examples
- Session 2
- Proper data example
- Surviving the Titanic
- Further resources
- Final comments

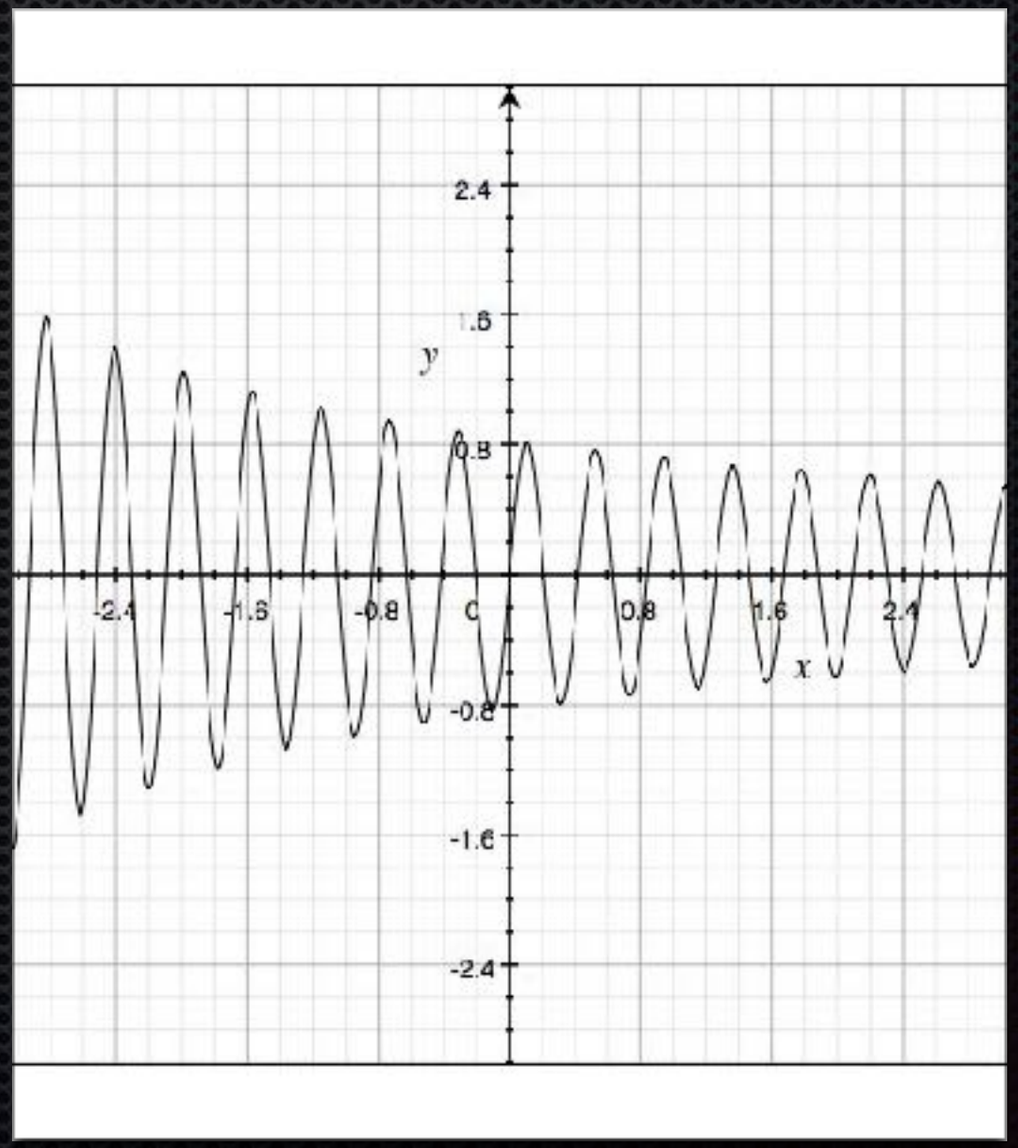
Introducing the problem

- Can we, given existing data make accurate predictions?
- Need correlation, 2 variables simple, after 3 hard to visualise.
- Figure shows example, can easily draw regions.



What is Machine Learning (ML)

- Finding boundaries in data (Curve fitting)
- Statistical tools assume Gaussians, linearity.
- Think optimisation problem, local vs global maximum.



- ML like a random walk.
- ML allow us to find new correlations, better distinctions.
- Computer resources are cheap.
- Want to base our limits on (changing) data.
- Almost a one solution for everything… At least changing data within reason.

What is Machine Learning (ML)

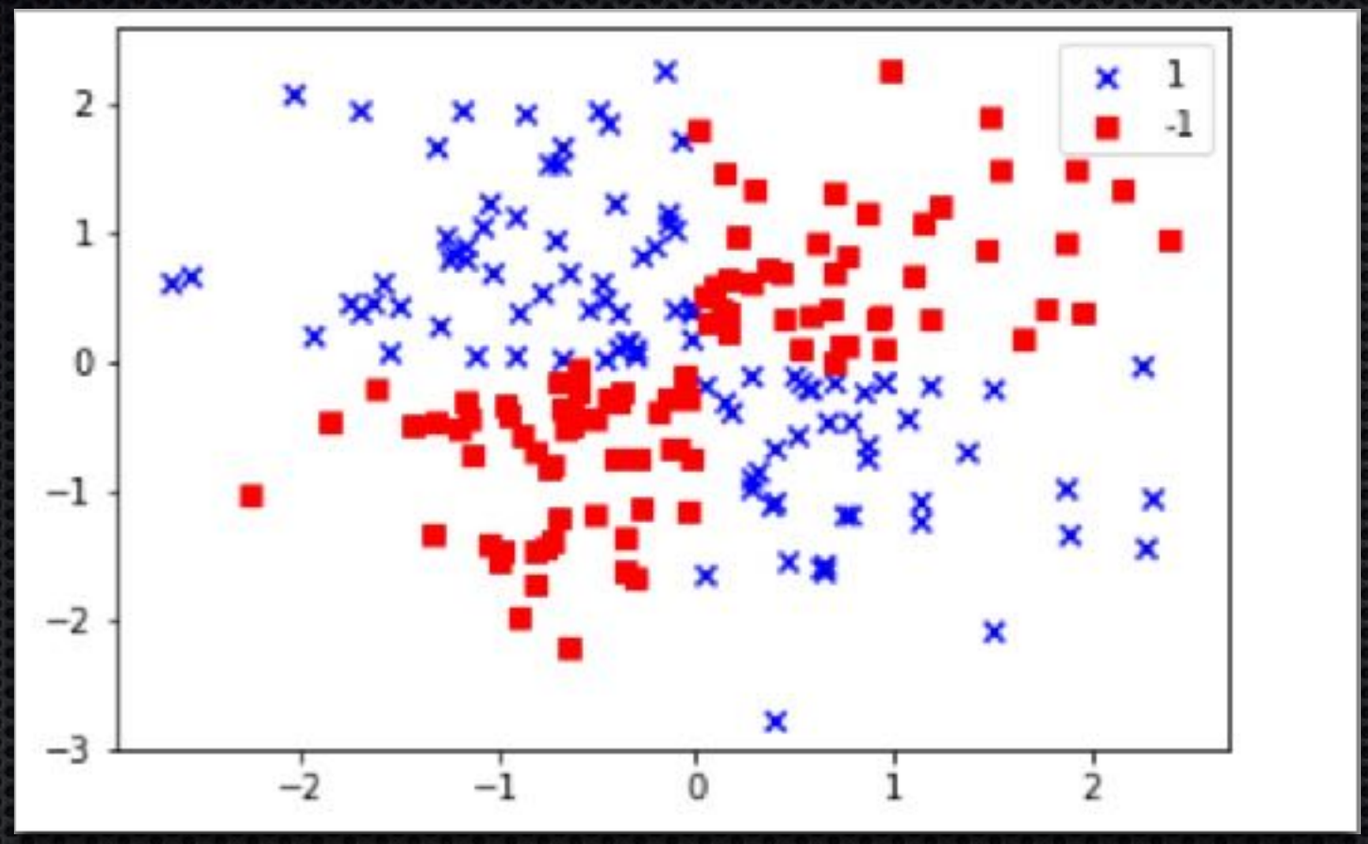
- We want an if(Boolean statement) where this statement is generated from the data

```
if(.........):
    retVal = 'car'
else:
    retVal = 'bike'
```

What is Machine Learning (ML)

- We want an if(Boolean statement) where this statement is generated from the data

```
if(object.wheels == 4):
    retVal = 'car'
else:
    retVal = 'bike'
```



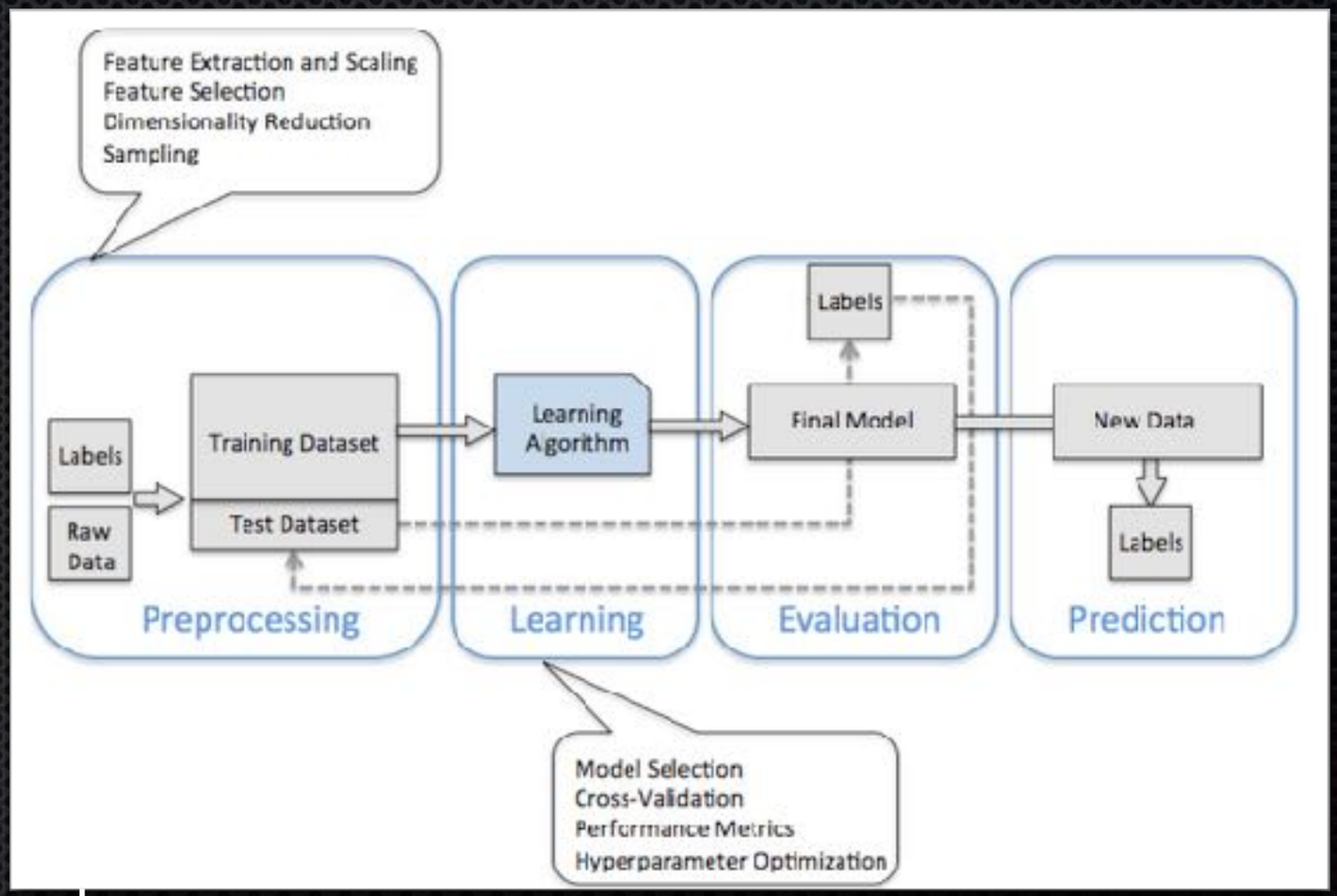
Translate this into boundaries

Never forget

(Non)-linear transforms to highlight differences in the parameters.

What is Machine Learning (ML)

- Splitting data into Training, Testing and “Real”.
- Training used for the algorithm
- Testing to evaluate
- “Real” is yet unclassified data



What is Machine Learning (ML)

- Example
- Everyone in here is in research.
- Could use us as a training set.
- Bad statistics, would find only one region.
- Would become apparent with test data.

What it is not

- Magical tool, still needs good data, good statistics
- Much of the work goes into understanding the data.
- May find regions and make predictions.

Never forget, Correlation -/> Causation

Can find “flukes”.
All black cats have tails, not all cats with tails are black.
Can also find a good variable but miss others in the causation
Crisps causing cancer, actually fried fat.

- We analyse data, thus we can only talk about the data.
- Depends on how the data was collected.
- At best implies a causation, or points towards a trend
  a reason to continue research.

Example of use

- Similar problems, all relevant for ML.
- Time series and forecasting
- Spatial data analysis
- Text analytics
- Image analysis
- Other

Time series and forecasting

- Finance
- Power usage/production
- General production of goods

Spatial data analysis

- Heavy metal concentration in water, building probability/concentration regions using some data points.
- Crime rate
- Geography of demographic

Text analytics

- Classify if the world is happy or sad
- Look at tweets, facebook. Can we estimate if the messages are happy or sad?
- Translation (Really cool, not just translate using data but make comparative language structures)

Image analysis

- Recognise a face? A sign? Handwriting?
- Security
- Self-driving cars
- Analysing handwritten text
- (Don’t forget to add pictures at night)

Other

- Predict a users preferences if he watches a tv-show and or movies (Amazon, Netflix)
- Will a flight be delayed? By how much? If only we had the data…
- Video classification
- Increase productivity, efficiency based on worker
- Improve health care

Example algorithm/method

- KNN, how very other, close points/objects/events classified?
- Trees, think normal statistical tree, if this and this and this then ->, of course uses a feedback loop to be created.
- Neural nets are similar to trees.
- Logistic regression, close to normal probability calculations.
- Etc

Big data

- Could we find correlation between heart disease and driving?
- Blood pressure (over normal) and Cancer?
- Buying Iphones and eating apples for lunch?
- Detector variables and particle type?

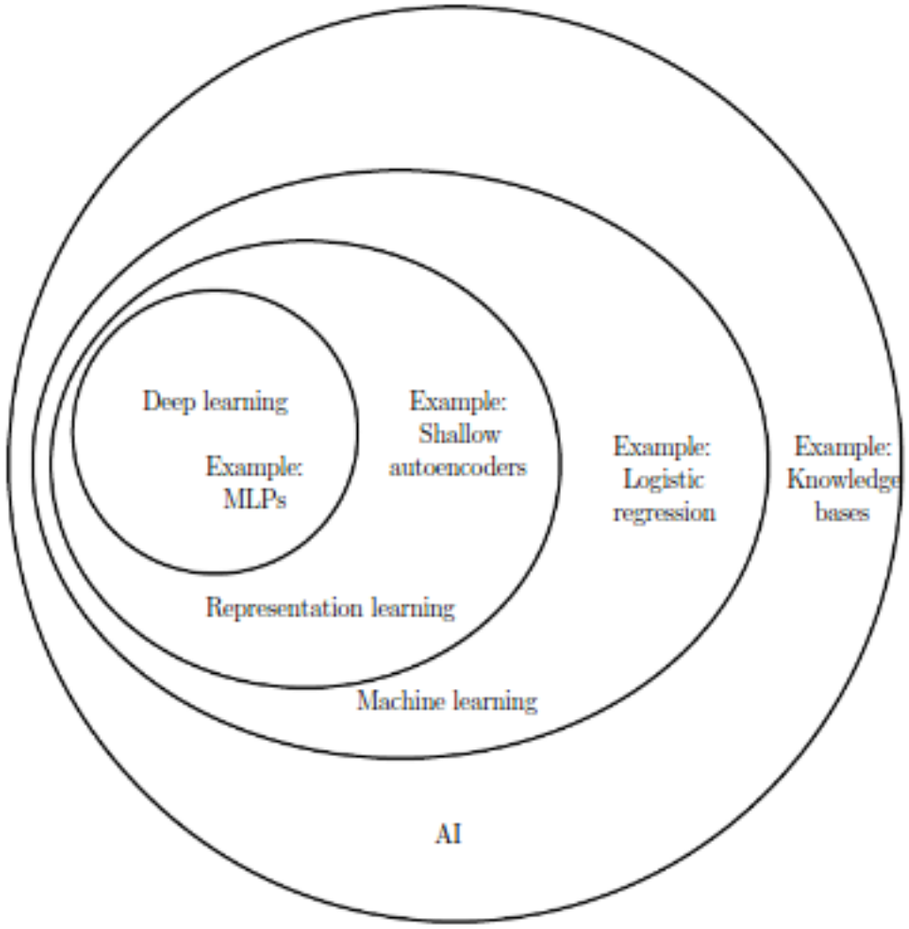
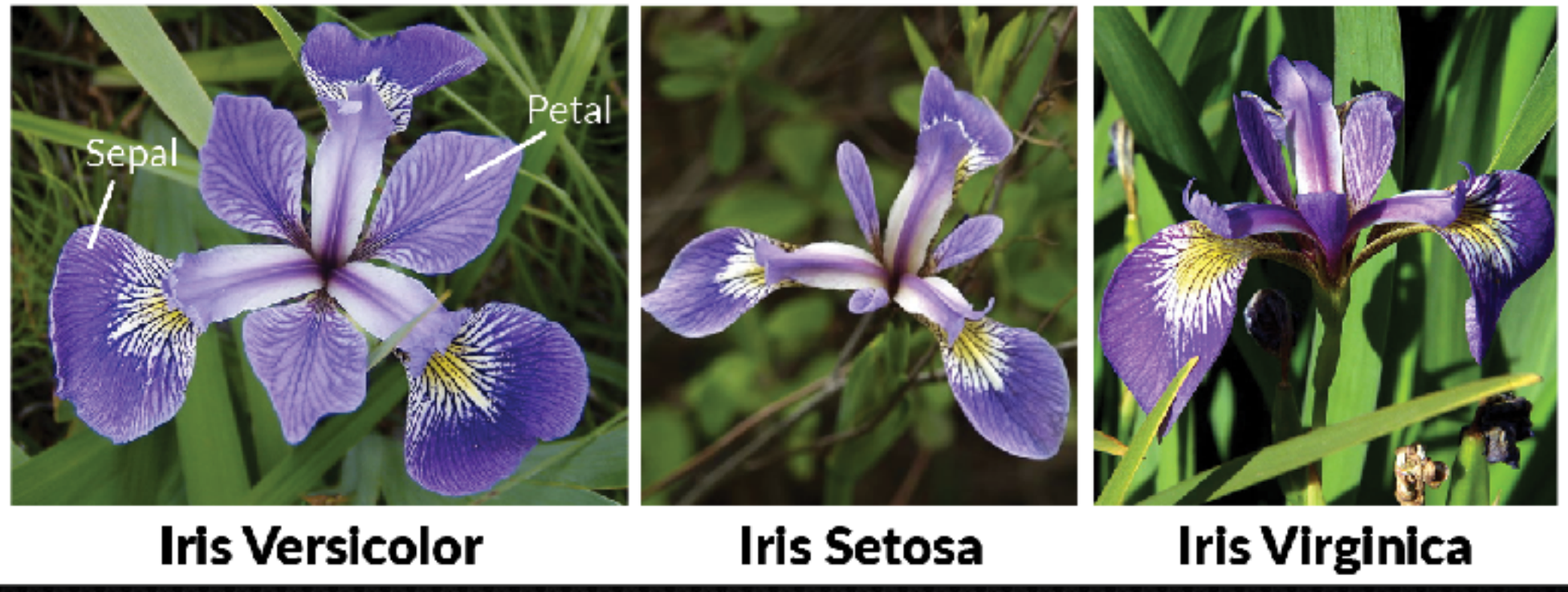


Figure 1.4: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.

ML vs Ai - http://www.deeplearningbook.org



Iris classification

Good initial problem

- Get you started with the tools
- Visualise what you are doing, the regions.
- First 2 different types
- Finishing with 3.

- Giving you free roaming with two different code snippets and two different algorithms, perceptron and adaptive linear neutron classifier

- Try getting into the code if possible, change parameters and see what the difference is.

- Google the code and find the documentation.

Post-analysis

- Hopefully you've had a change to play around.
- One thing we didn't do is preprocessing the data, i.e using different (non)linear transforms to highlight differences in the parameters.

- When would you want fixed, hard line boundaries?
- What would be the advantage?
- What about soft boundaries?
- What could go wrong?

Surviving the Titanic

https://www.kaggle.com/c/titanic/data

A data analysis competition

- Start of by understanding our data.
- Try finding important features.
- Simplify data, don't oversaturate the number of variables
- Classifications should (often) be numerical
- Test different ML models, (the easy and quick thing)

Post-analysis

- Lets not make the obvious mistakes
- We can not say:
- Women have a higher chance of survival in cold water
- Higher class -> Richer -> Higher chance of survival in cold water.
- Missing the context!

Reading handwritten numbers in Tensorflow

- Many many resources online and official tutorials.
- Using the following example
- https://github.com/aymericdamien/TensorFlow-Examples/

Further resources

- http://www.ppe.gla.ac.uk/~phallsjo/files/CardiffSTFC/
- https://www.edx.org/course/applied-machine-learning-microsoft-dat203-3x-3
- https://www.edx.org/course/principles-machine-learning-microsoft-dat203-2x-5
- http://tmva.sourceforge.net - Particle physics
- https://www.kaggle.com
- Python Machine Learning - Sebastian Raschka
- http://www.deeplearningbook.org - Ian Goodfellow Yoshua Bengio Aaron Courville

Final comments

- A brief introduction
- Many more challenges online
- Many packages with tutorials and examples
- Understand and handle the data
- ML algorithms are important
- Understanding and filtering the data is more important

Thank you for attending

Any questions?

Otherwise just approach me at any time or send me an email.
p.hallsjo.1@research.gla.ac.uk