What is Machine Learning?

Dr. Johan Hagelbäck

johan.hagelback@lnu.se

http://aiguy.org

What is Machine Learning?

”The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.”

Tom Mitchell

- … or more formally:

What is Machine Learning?

” 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

- If we take marking email as spam/non-spam as example:
  - **E** is the emails we have collected
  - **T** is the task of classifying an email as spam or non-spam
  - **P** is a performance measure, for example how many of our emails that are correctly classified (100% as best, 0% as worst)

What is Machine Learning?

- How can we classify email as spam/non-spam?
  - Find common patterns (e.g. words like ”cheap Rolex”) in spam emails
  - Write rules:

    if email contains “cheap Rolex” then mark as Spam

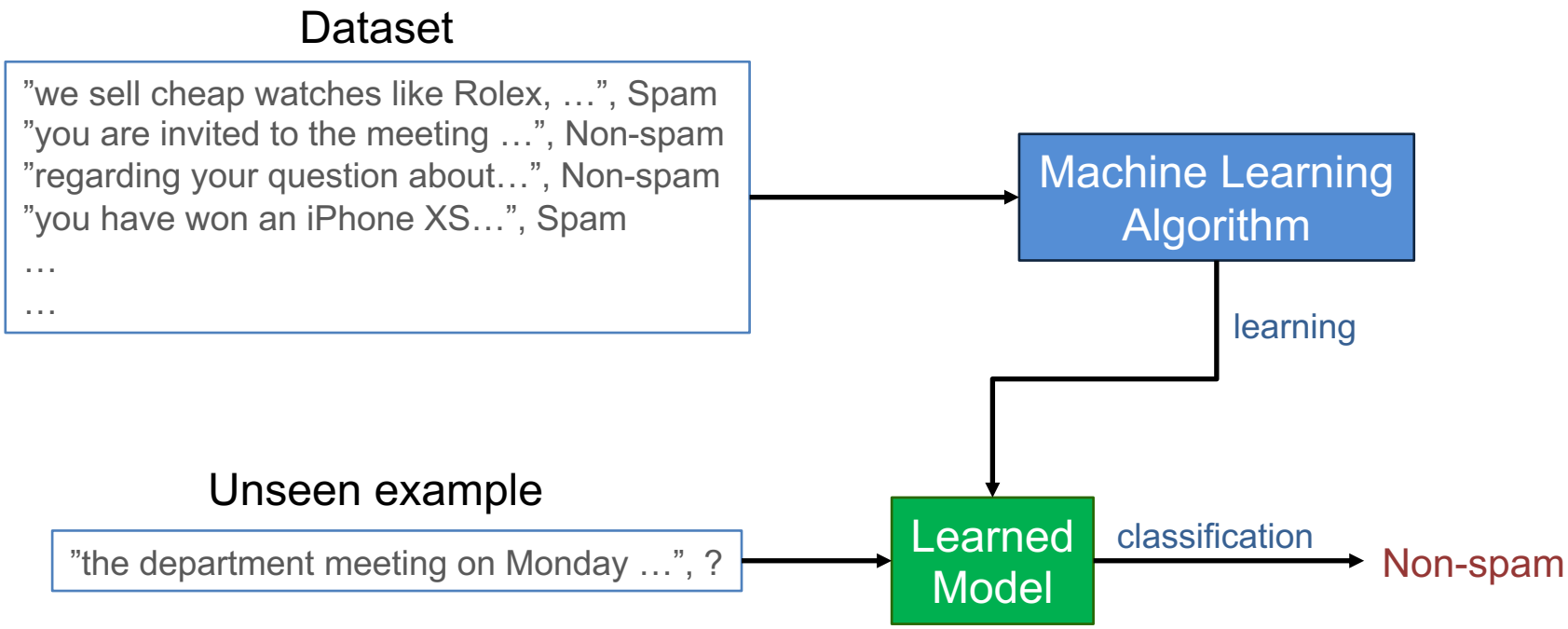
  - Update and maintain rules as new types of spam emails are received

- Problems:
  - Maintaining the rule base is time consuming
  - It can be difficult to find general patterns that include all spam and exclude all non-spam

What is Machine Learning?

- A better approach:
  - Collect a dataset of example emails
  - Manually label all emails as spam or non-spam
  - Construct a computer program that can learn from data
  - Use the program to learn the task of email spam/non-spam classification
  - Use the learned program to classify new emails when they are received
- Machine Learning is the construction of programs that learns from data without being explicitly programmed
- Instead of the programmer telling the program how to find spam emails, the program learns from the data how to do it

What is Machine Learning?



When do we need ML?

- A traditional program for solving some problem has some properties like:
  - You know or can control the inputs to the program
  - You can specify how the program will achieve its goal
  - You can map out what decisions the program will make and under what conditions it makes them
  - You can test your program with known inputs and be confident that it works correctly

When do we need ML?

- For some problems you cannot write a traditional program to solve it
- They have properties such as:
  - The scope of all possible inputs is not known beforehand
  - You cannot specify how to achieve the goal of the program, only what that goal is
  - You cannot map out all the decisions the program will need to make to achieve its goal
- Here, machine learning comes to our rescue!

When do we need ML?

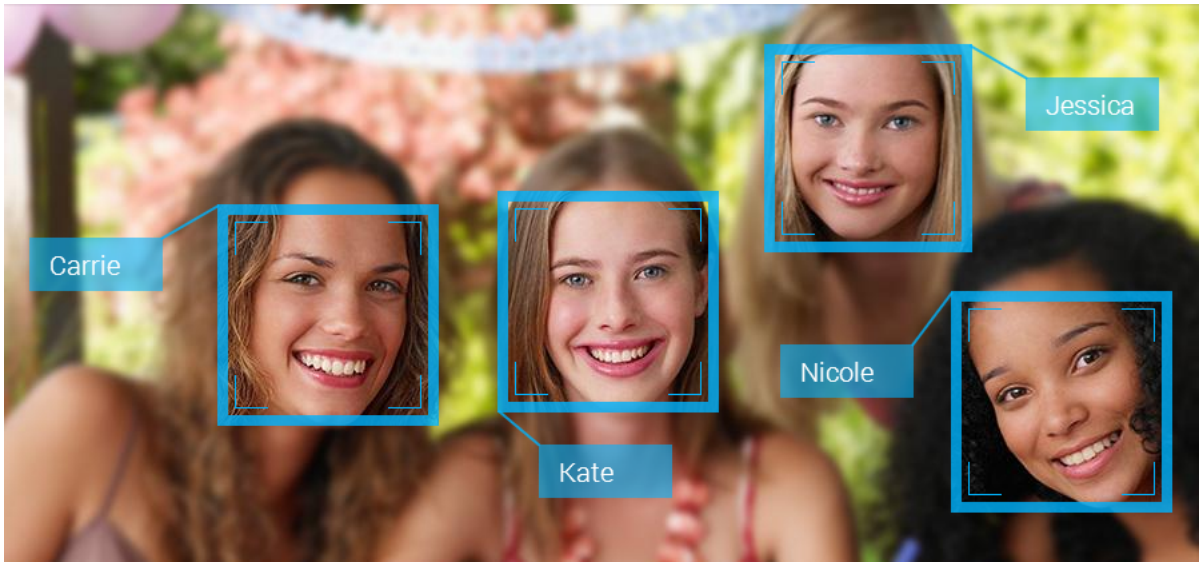
- Machine Learning can be used for complex problems where:
  - We might not know exactly how inputs look like
    - How many different cat images can be found?
  - We have a goal, but it is difficult to understand *how* to achieve it and what decisions the program have to make to achieve it
    - How can we write rules to detect a cat?
  - We can however collect and label data
    - Collect pictures with and without cats and label them as "Cat" and "No cat"

Benefits of ML

- Automatically:
  - Machine Learning algorithms use data to create models that knows how to achieve the goal
- Fast:
  - ML algorithms can analyze and find solutions faster than you can manually program a solution
- Accurate:
  - ML algorithms can run longer and on more data than you can, thus creating more accurate solutions
- Scale:
  - ML algorithms scale well to big data, which Excel sheets don’t

Examples of ML

- Finding and recognizing faces:



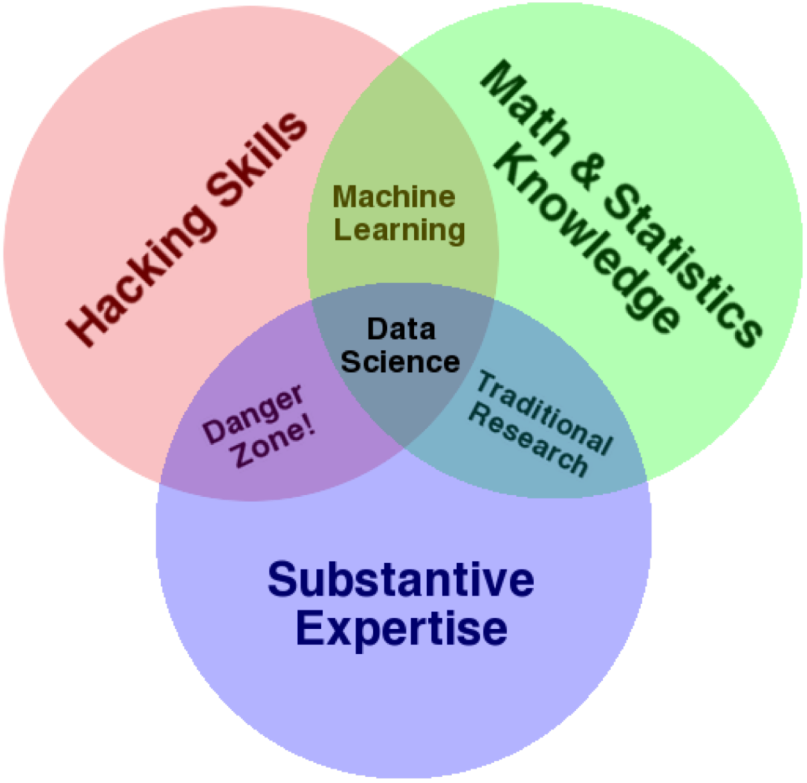
Examples of ML

- Product Recommendations:
  - Recommending movies on Netflix or songs on Spotify based on what you previously have consumed
- Medical Diagnosis:
  - ML systems are today equal or better than human experts in detecting skin cancer
- Recognizing handwritten letters and digits:
  - Automatically read addresses on physical letters

When can ML not be used?

- Inconsistent data:
  - One person can label the same movie as ”thriller”, another as ”action thriller”, a third as ”action movie”, …
  - Minor inconsistencies are ok, but severe can cause problems
- No mapping between input and labels:
  - There must exist a mapping between the inputs and labels which the ML algorithm can find
  - For example the color of a car doesn’t map to engine power (unless we forbid all brands to use red except Ferrari…)
- Difficulties in collecting and/or labelling data:
  - We must be able to collect enough data and in some way label it
  - Can be difficult and time-consuming to label images of *skin cancer* and *not skin cancer* (verification using lab tests)

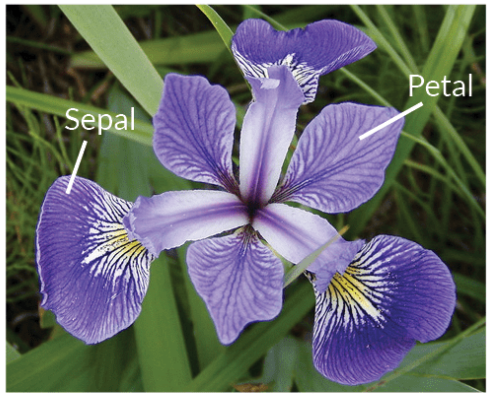
Where do ML fit in?



Venn diagram by Drew Conway

Common example

- Classify an iris flower into one of the three species
- Inputs are measurements of petal length and width, and sepal length and width:



Iris Versicolor

Iris Setosa

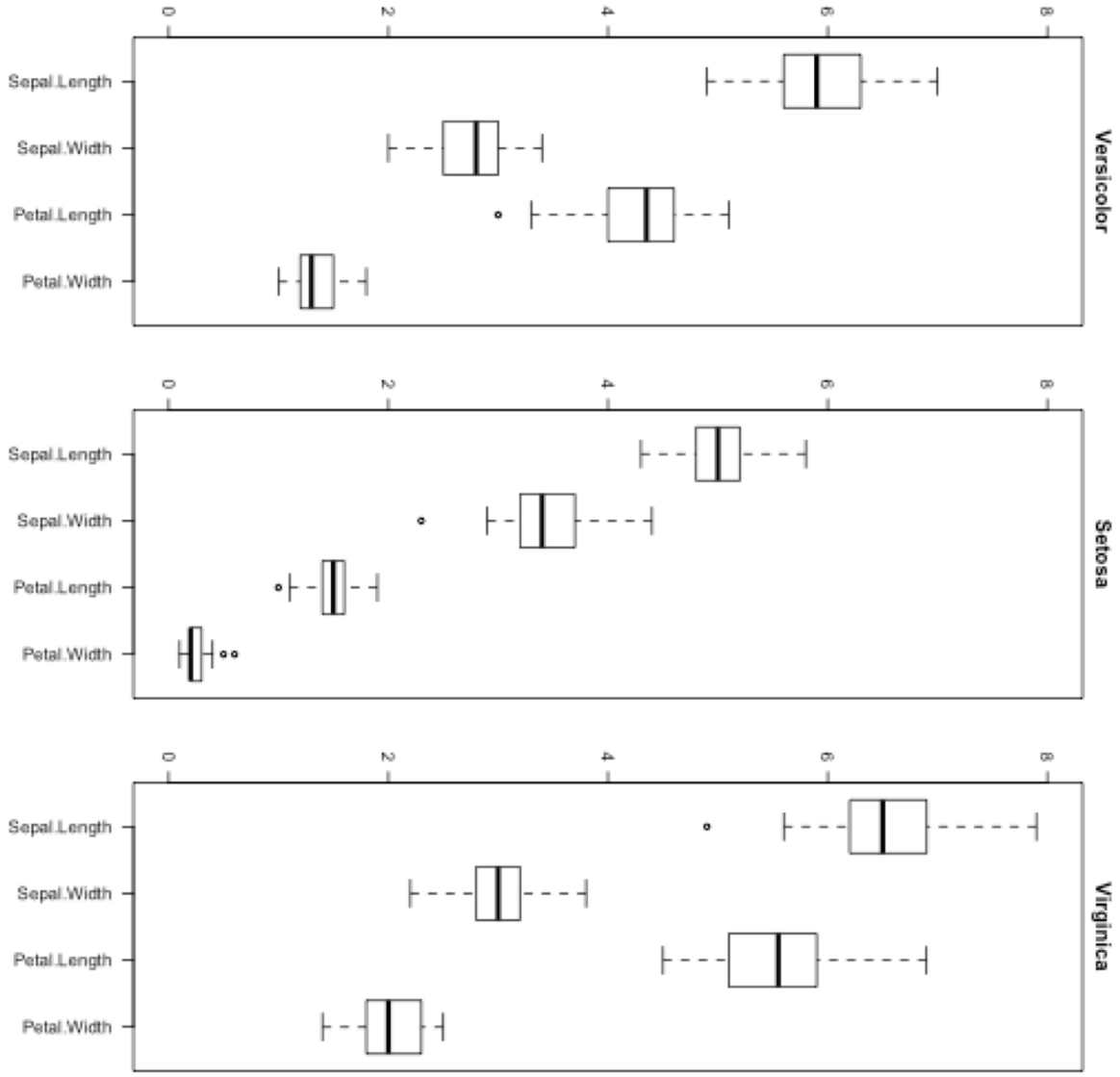
Iris Virginica

Iris dataset

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	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

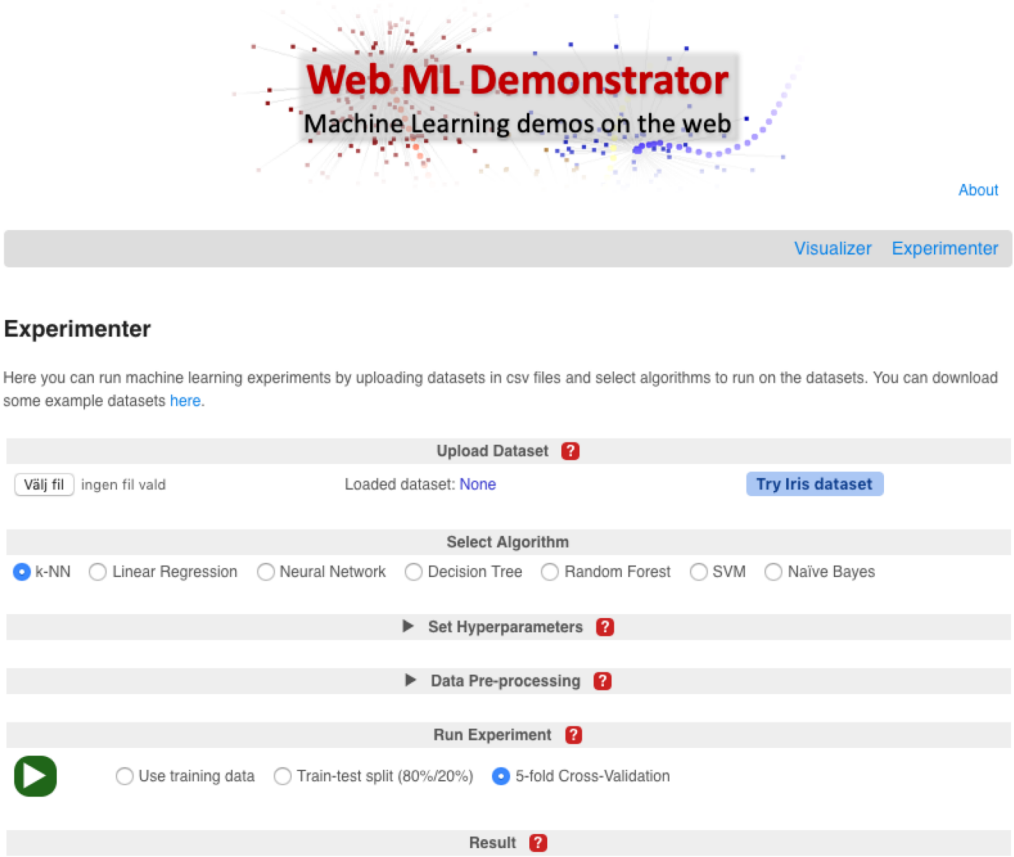
- A random guess would result in 33% accuracy (correct classifications)
- Can we do better?

Classification – traditional approach



In a traditional program we write rules for discrimination between sepal and petal lengths and widths!

Classification – machine learning



aiguy.org/webml/experimenter.html

What is Machine Learning?

Dr. Johan Hagelbäck

johan.hagelback@lnu.se

http://aiguy.org