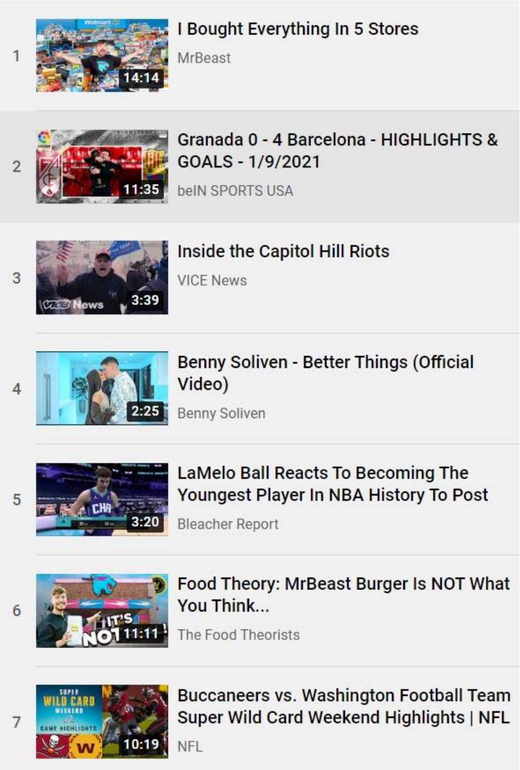
IN3050/4050, Lecture 1
Introduction to Machine Learning

In this video, we will give a short introduction to the field of machine learning, covering the first chapter of the Marsland textbook. The terminology and techniques discussed today will be with us for the rest of the course.

How many products have you used in the previous week that relies on machine learning? Take a few seconds and think it over. The answer is probably a lot. Social media like Facebook and Twitter, content pages like youtube, spotify and netflix, search engines like google, even stores like komplett or elkjøp use machine learning to get insights into their users, increase relevance and provide a more personalized service.



1 I Bought Everything In 5 Stores
MrBeast
14:14

2 Granada 0 - 4 Barcelona - HIGHLIGHTS & GOALS - 1/9/2021
beIN SPORTS USA
11:35

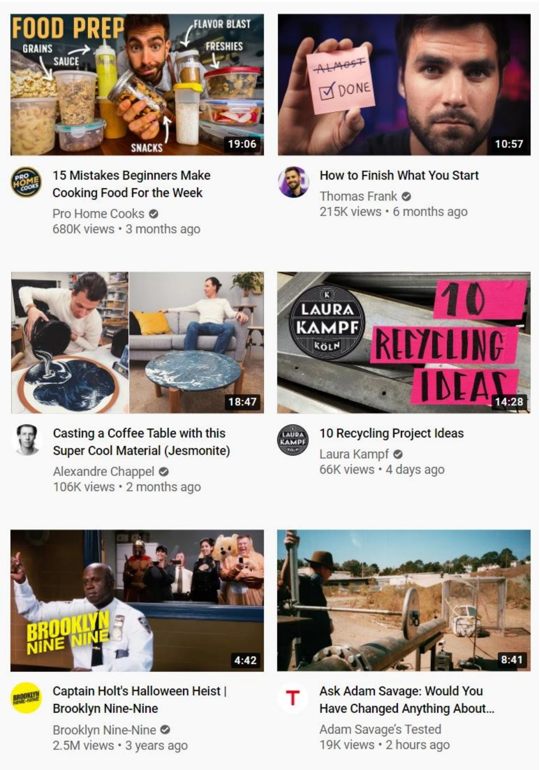
3 Inside the Capitol Hill Riots
VICE News
3:39

4 Benny Soliven - Better Things (Official Video)
Benny Soliven
2:25

5 LaMelo Ball Reacts To Becoming The Youngest Player In NBA History To Post
Bleacher Report
3:20

6 Food Theory: MrBeast Burger Is NOT What You Think...
The Food Theorists
11:11

7 Buccaneers vs. Washington Football Team Super Wild Card Weekend Highlights | NFL
NFL
10:19



15 Mistakes Beginners Make Cooking Food For the Week
Pro Home Cooks
680K views • 3 months ago
19:06

How to Finish What You Start
Thomas Frank
215K views • 6 months ago
10:57

Casting a Coffee Table with this Super Cool Material (Jesmonite)
Alexandre Chappel
106K views • 2 months ago
18:47

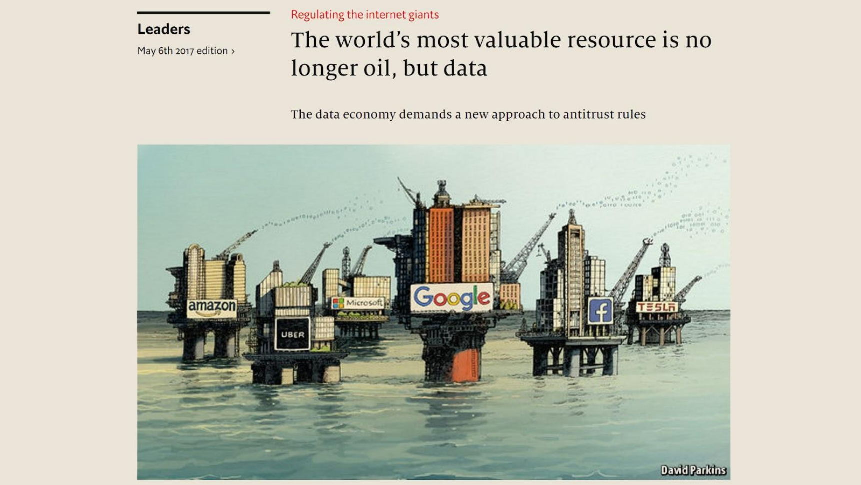
10 Recycling Project Ideas
Laura Kampf
66K views • 4 days ago
14:28

Captain Holt's Halloween Heist | Brooklyn Nine-Nine
Brooklyn Nine-Nine
2.5M views • 3 years ago
4:42

Ask Adam Savage: Would You Have Changed Anything About...
Adam Savage's Tested
19K views • 2 hours ago
8:41

Here, we see the most popular youtube videos as of january 2021 (left image). If I go to the front page of youtube, I am instead presented with this selection of videos (right image). I went through the 200 most popular videos, and none of these appear on my suggested videos. That is because Youtube's machine learning algorithms deem these to be more relevant for me, with a higher probability for me to watch them. Looking at the current top 200, I think they are right.

Youtube is able to give this personalized experience due to the massive amount of data they have on each user. The whole world is now driven by data, and more and more is available every day. Over 500 hours of video is uploaded to youtube every minute. That’s 30.000 hours of uploaded content every hour! The large hadron collider in Switzerland produces over 60 gigabytes per minute, and over a billion credit card transactions are done every single day. These massive amounts of data are impossible for humans to work with directly, but machine learning allows us to use the data for something valuable.



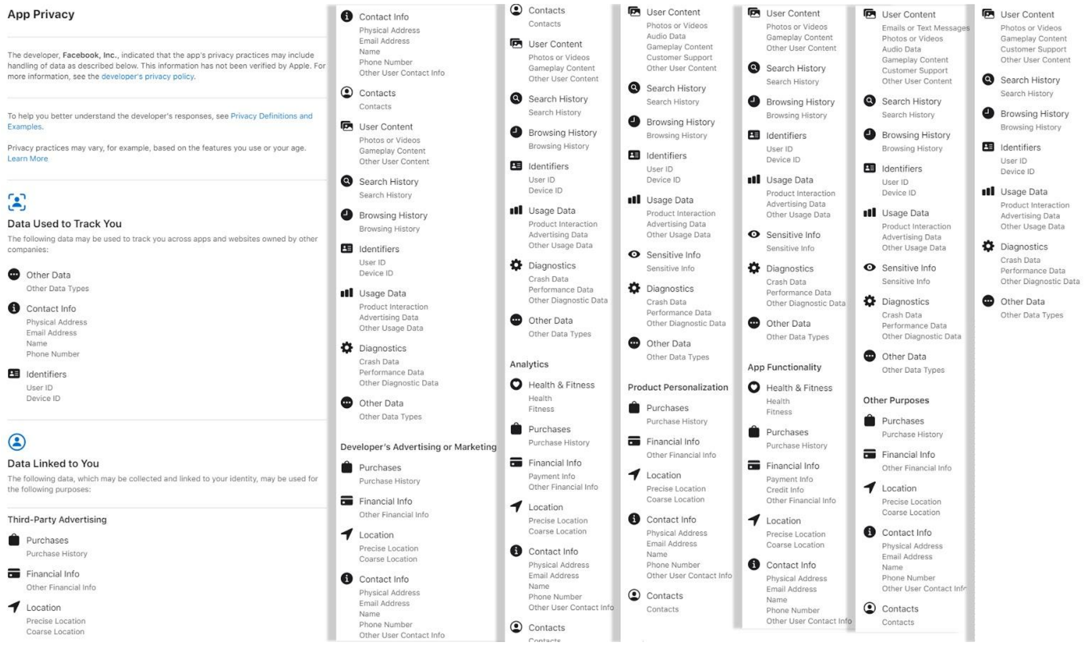
Leaders
May 6th 2017 edition ›

Regulating the internet giants

The world’s most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

This article from the Economist says the world’s most valuable resource is no longer oil, but data. Four out of the five most valuable companies in the world in 2020 were technology companies. Microsoft, Apple, Amazon and Google are all heavily dependent on data and machine learning.



App Privacy

The developer, **Facebook, Inc.**, indicated that the app's privacy practices may include handling of data as described below. This information has not been verified by Apple. For more information, see the developer's privacy policy.

To help you better understand the developer's responses, see Privacy Definitions and Examples.

Privacy practices may vary, for example, based on the features you use or your age. Learn More

Data Used to Track You

The following data may be used to track you across apps and websites owned by other companies:

Other Data
Other Data Types

Contact Info
Physical Address
Email Address
Name
Phone Number

Identifiers
User ID
Device ID

Data Linked to You

The following data, which may be collected and linked to your identity, may be used for the following purposes:

Third-Party Advertising

Purchases
Purchase History

Financial Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

Contacts
Contacts

User Content
Photos or Videos
Gameplay Content
Other User Content

Search History
Search History

Browsing History
Browsing History

Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

Developer’s Advertising or Marketing

Purchases
Purchase History

Financial Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

Contacts
Contacts

User Content
Photos or Videos
Gameplay Content
Other User Content

Search History
Search History

Browsing History
Browsing History

Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

Analytics

Health & Fitness
Health
Fitness

Purchases
Purchase History

Financial Info
Payment Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

Contacts
Contacts

User Content
Photos or Videos
Audio Data
Gameplay Content
Customer Support
Other User Content

Search History
Search History

Browsing History
Browsing History

Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

Sensitive Info
Sensitive Info

Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

Product Personalization

Purchases
Purchase History

Financial Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

Contacts
Contacts

User Content
Photos or Videos
Gameplay Content
Other User Content

Search History
Search History

Browsing History
Browsing History

Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

Sensitive Info
Sensitive Info

Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

App Functionality

Health & Fitness
Health
Fitness

Purchases
Purchase History

Financial Info
Payment Info
Credit Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

User Content
Emails or Text Messages
Photos or Videos
Audio Data
Gameplay Content
Customer Support
Other User Content

Search History
Search History

Browsing History
Browsing History

Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

Sensitive Info
Sensitive Info

Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

Other Purposes

Purchases
Purchase History

Financial Info
Other Financial Info

Location
Precise Location
Coarse Location

Contact Info
Physical Address
Email Address
Name
Phone Number
Other User Contact Info

Contacts
Contacts

User Content
Photos or Videos
Gameplay Content
Customer Support
Other User Content

Search History
Search History

Browsing History
Browsing History

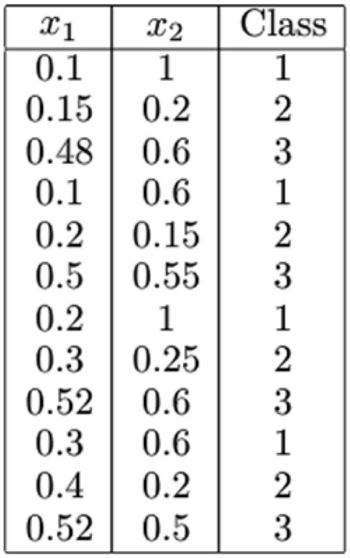
Identifiers
User ID
Device ID

Usage Data
Product Interaction
Advertising Data
Other Usage Data

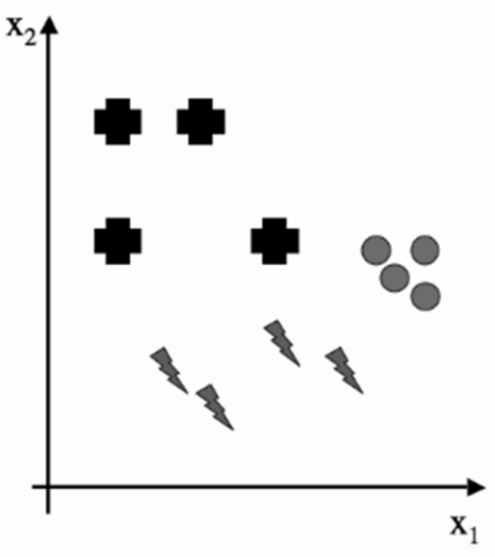
Diagnostics
Crash Data
Performance Data
Other Diagnostic Data

Other Data
Other Data Types

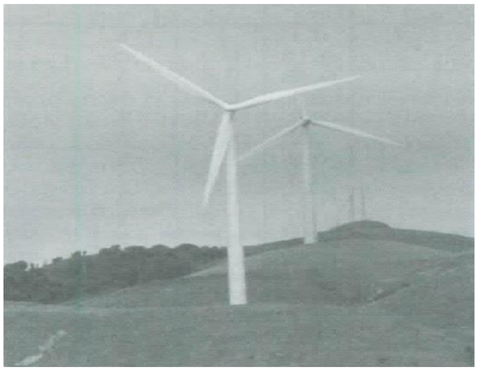
Facebook is the next on the list, and is definitely among the most aggressive in data collection. If we look at the newly required privacy notice for the facebook app in Apple’s App Store, we see just some of the information facebook collects about its users. This includes search and browsing history, user content like photos and video, location and financial information. This information is used by facebook to improve their own services, increase profitability of their ads, as well as selling and sharing with others.

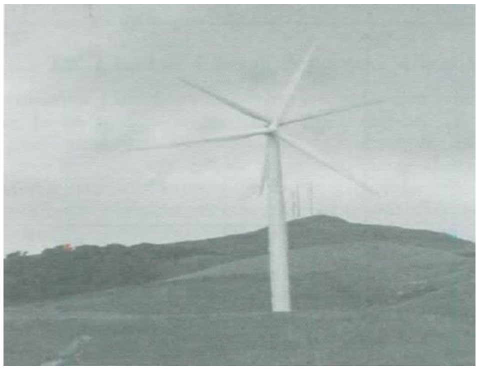


x_1	x_2	Class
0.1	1	1
0.15	0.2	2
0.48	0.6	3
0.1	0.6	1
0.2	0.15	2
0.5	0.55	3
0.2	1	1
0.3	0.25	2
0.52	0.6	3
0.3	0.6	1
0.4	0.2	2
0.52	0.5	3



Humans are generally pretty bad at interpreting even fairly simple data. Looking at this table makes little sense to us, other than the fact that there are three classes with two dimensional features. If we visualize the data in two dimensions, we start to see the structure or patterns in the data. This could give us some useful insight, however this is only possible for two or three dimensional data. Most interesting problems include more than this, along with many more data points than what we saw in this example.





There are many techniques to reduce the number of dimensions. They all, however, end up hiding information that might mislead us to draw the wrong conclusions. Here, we see two windmills. They are, of course, three dimensional in the real world, but are reduced to two dimensions in this picture. If we take the same picture from another angle, it looks like we have only one windmill with six blades instead. The same happens for more complex datasets. By reducing the dimensionality, we could lose important information as well. That being said, many techniques that do this can be a powerful tool, which we'll come back to in a later lecture.

So what is machine learning? The book defines machine learning as “the ability of a program to learn from experience” — that is, to modify its execution on the basis of newly acquired information. The program does not do exactly what it has been programmed to do, but uses its experience to affect its actions. Machine learning is about automatically extracting relevant information from the data it has been given, and then applying it to analyze new data it encounters.

A very important issue in machine learning, which we will get back to many times, is the principle of generalization. We want our machine learning techniques to learn something general -- something that works not only for the data from the training set, but also on previously unseen data. If we train a self-driving car in Oslo, we want that car to also work in Drammen, Stavanger or Narvik. We want it to learn about driving in general, not specifically driving in Oslo. This is typical for most machine learning. We want it to learn how to solve the problem for new examples as well, not just the ones it encountered during learning. Humans are really good at this. We are able to apply knowledge we learn to new situations we haven’t been in before. Hopefully you’ll all be able to go work in industry and academia, and use the knowledge you’ve gained in this course to solve a wide range of challenges going way beyond the simple problems you tackle at the university.

So when should we use machine learning? The book suggests a number of situations where machine learning could be useful.

- Sometimes, human expertise does not exist. No humans are experts at walking and navigating on mars, so a robot able to learn and adapt while operating on mars itself could be more useful than relying on purely pre-programmed behavior.
- Humans might be unable to explain their expertise. We are experts at speech recognition, but explaining how this is done and implementing something similar is very challenging.
- Many things change over time, and allowing a system to change with its environment might allow systems like self driving cars to work in different areas and conditions.
- We’ve already talked about personalization. Netflix can’t hire someone to give all of us recommendations on what to watch, but they can build a machine learning system that learns based on their users' history and continuously improve its recommendations.

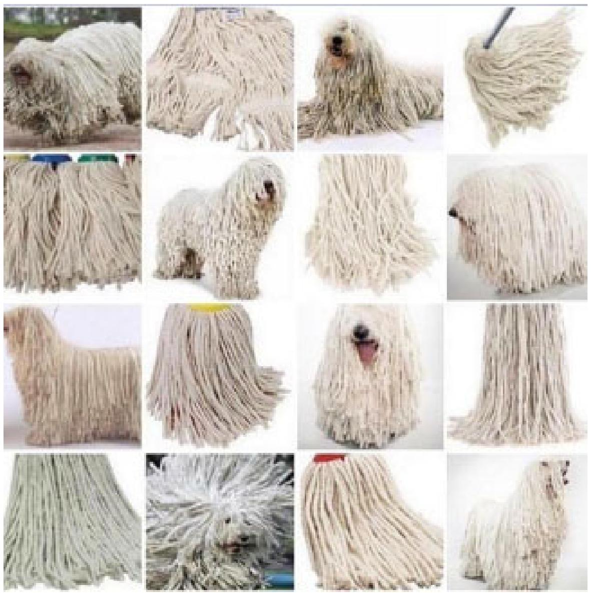
The book suggests formalizing the learning problem a bit to divide it into three distinct aspects. Learning is defined as improving with experience at some task. We have a task **T** that we wish to improve on. We measure how well we solve that task with a performance measure **P**. The sum of our experience, **E**, helps us improve our performance over time.

Let us look at a couple of examples. The game of checkers is a fairly simple board game. The task **T** would be to play the game. To train a machine learning agent, we need some way of evaluating the performance to guide the learning, **P**. There are many ways to do this, but one example could be the percentage of games won against an arbitrary opponent. It’s experience **E** might be in playing practice games against itself. As we’ll come back to in a later lecture, there are several downfalls to this specific approach to solving checkers, but it nicely illustrates this formalized way of analyzing the learning task.

Let’s take a look at another example. Imagine that we might want to make a system to sort mail with hand-written addresses. The task **T** would be to recognize hand-written words on envelopes. Performance **P** could be measured by the percentage of correctly classified words. Experience **E** could be based on hand-labeled images of handwritten letters.

There are three main types of machine learning. **Supervised**, **unsupervised** and **reinforcement learning**. These vary according to a couple of vital questions. How does the algorithm know whether it is getting better or not? And how does it know how to improve?

Supervised learning is, perhaps, the most straight-forward type. Here, training data includes the desired outputs, so the algorithm has examples of both the problem and the answers. Imagine that we want to be able to classify whether something is a dog or not. For supervised learning, our dataset would contain the label “dog” for all images of a dog. This might require a human to hand-label a large number of images, which could be very labor intensive. Supervised learning is generally simple to use and achieves good results, but the requirements to label the data limits the number of applicable problems.



In **unsupervised learning**, the training data does not include the desired outputs. Instead, the algorithm tries to identify similarities between the inputs so that similar items are categorized together. Feeding the dataset to the right into an unsupervised learning algorithm and specifying that you want two classes would split the dataset into two distinct groups, but is almost guaranteed to not give us “dog” and “not dog”, in this case a mop as the labels. Instead, you might get all the images with a completely white background in one group, with the rest in another.

In **reinforcement learning**, the algorithm is only told when the answer is wrong, but does not get told how to correct it. This is the type of learning we often use to teach a dog a new trick. If it does it correctly, we give it a treat. In the same way, a machine learning algorithm gets a signal that indicates whether it solved the problem correctly or not, but it has to figure out how to change its own behavior to get this reward more frequently. This is for instance very useful for playing games. In chess, we rarely know the perfect move to make throughout a game. Training an agent instead on, for instance, avoiding to lose pieces will allow it to improve its game playing ability over time. Many real world problems follow this structure where we don’t know the optimal choice, but we do get some feedback of performance during or after the task has been solved.

In this video, we have talked about the topic of machine learning, and briefly introduced supervised, unsupervised and reinforcement learning. More information is available in the first chapter of the Marsland textbook.