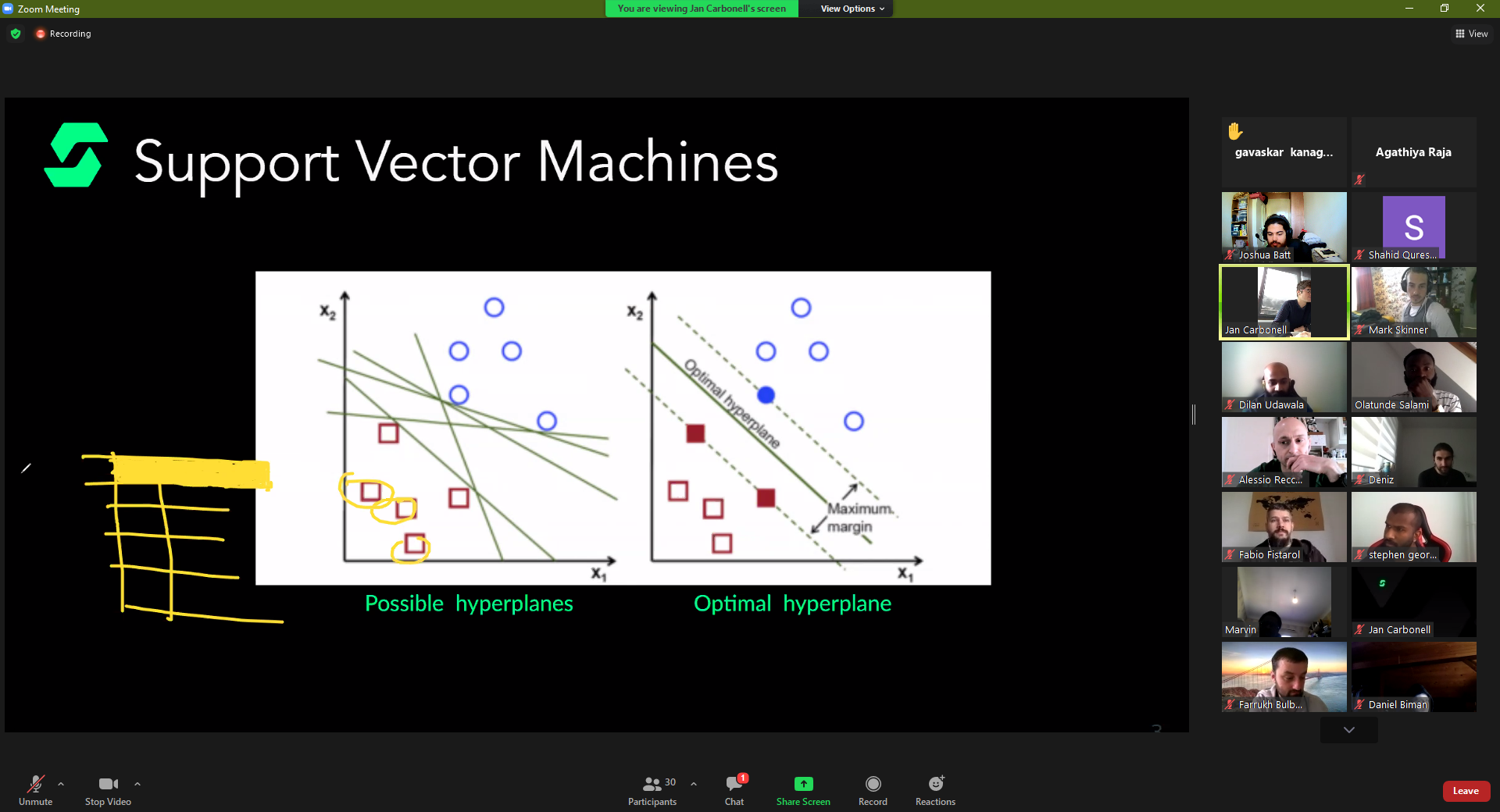
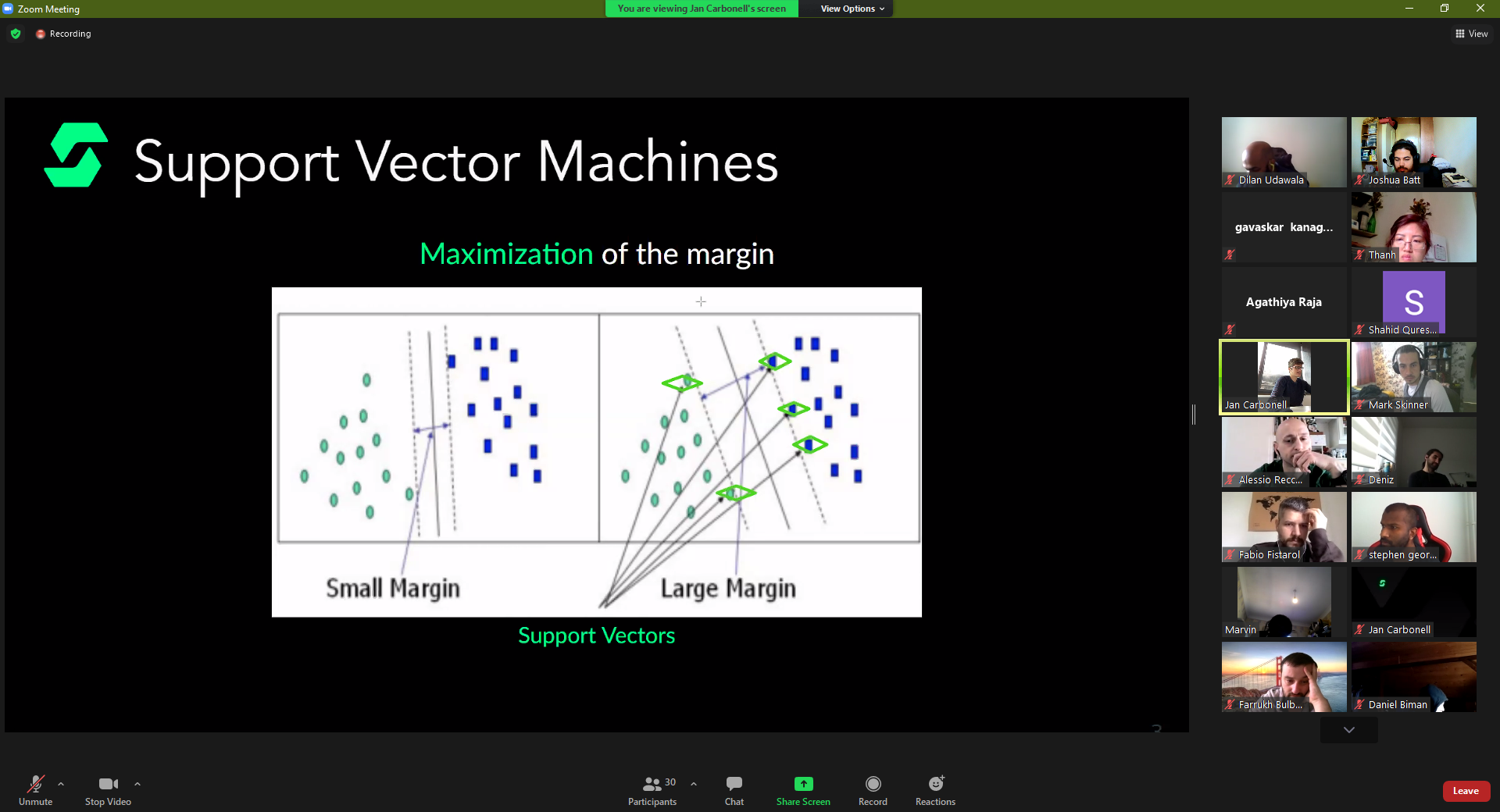
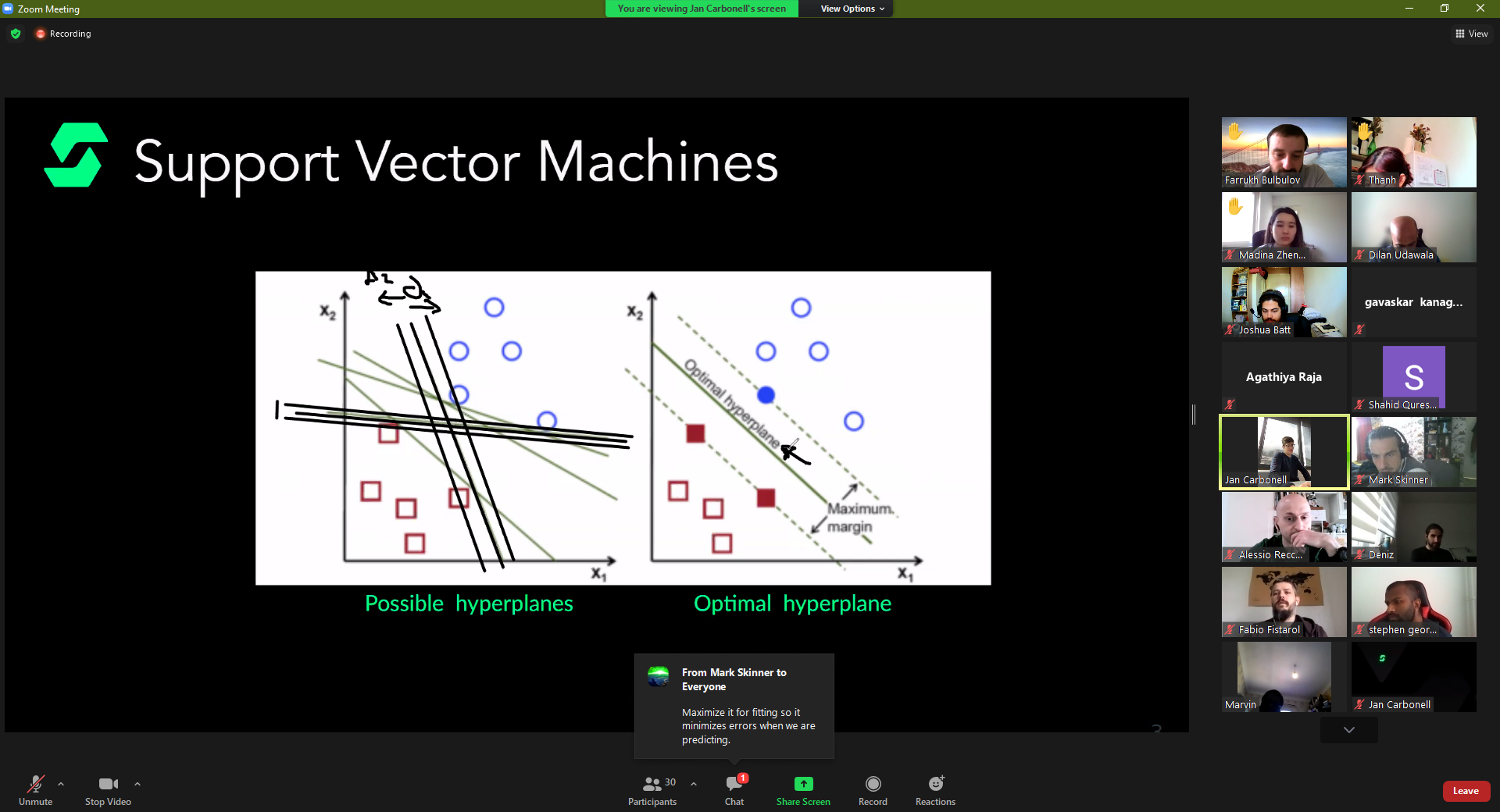
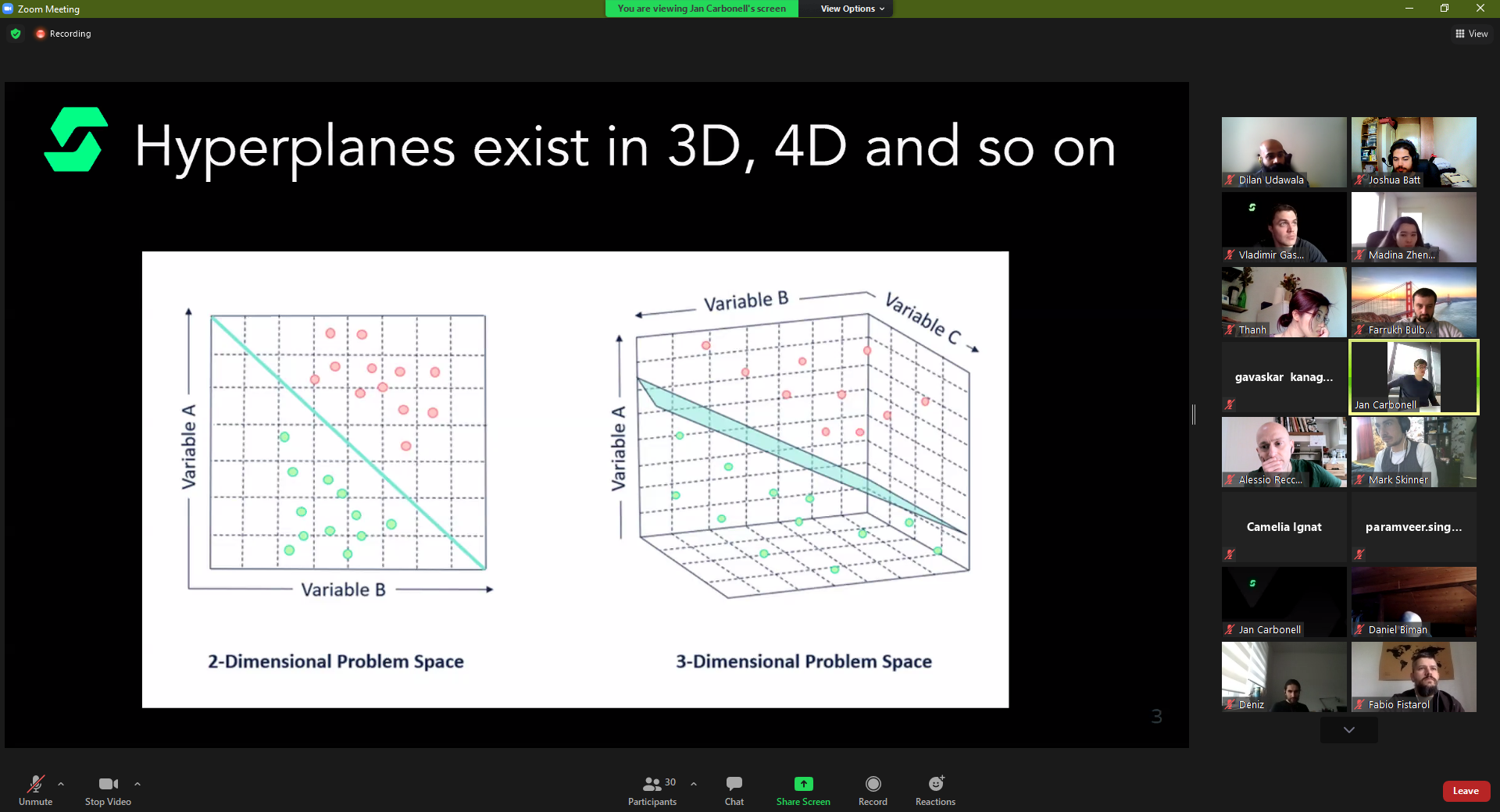
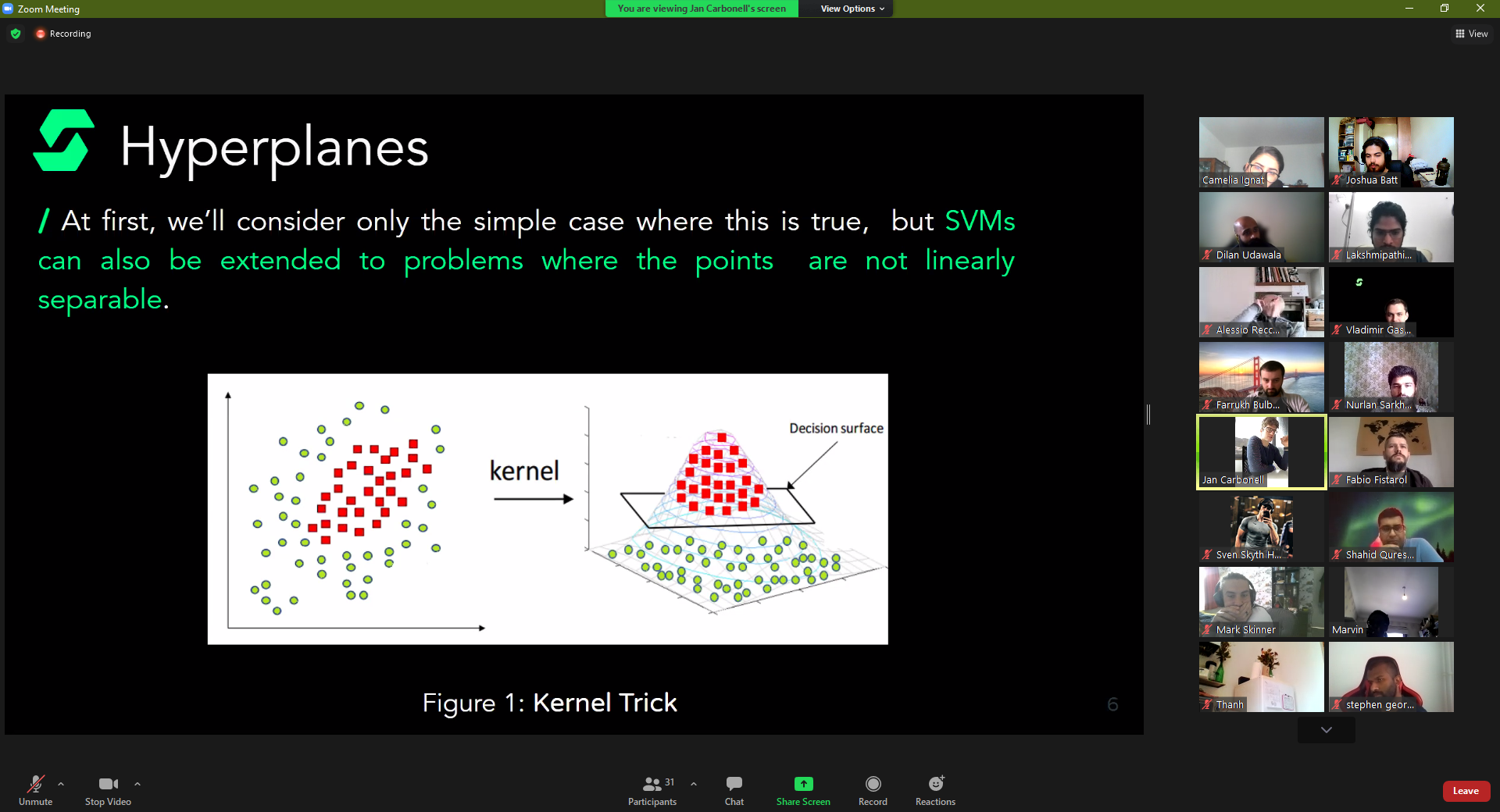
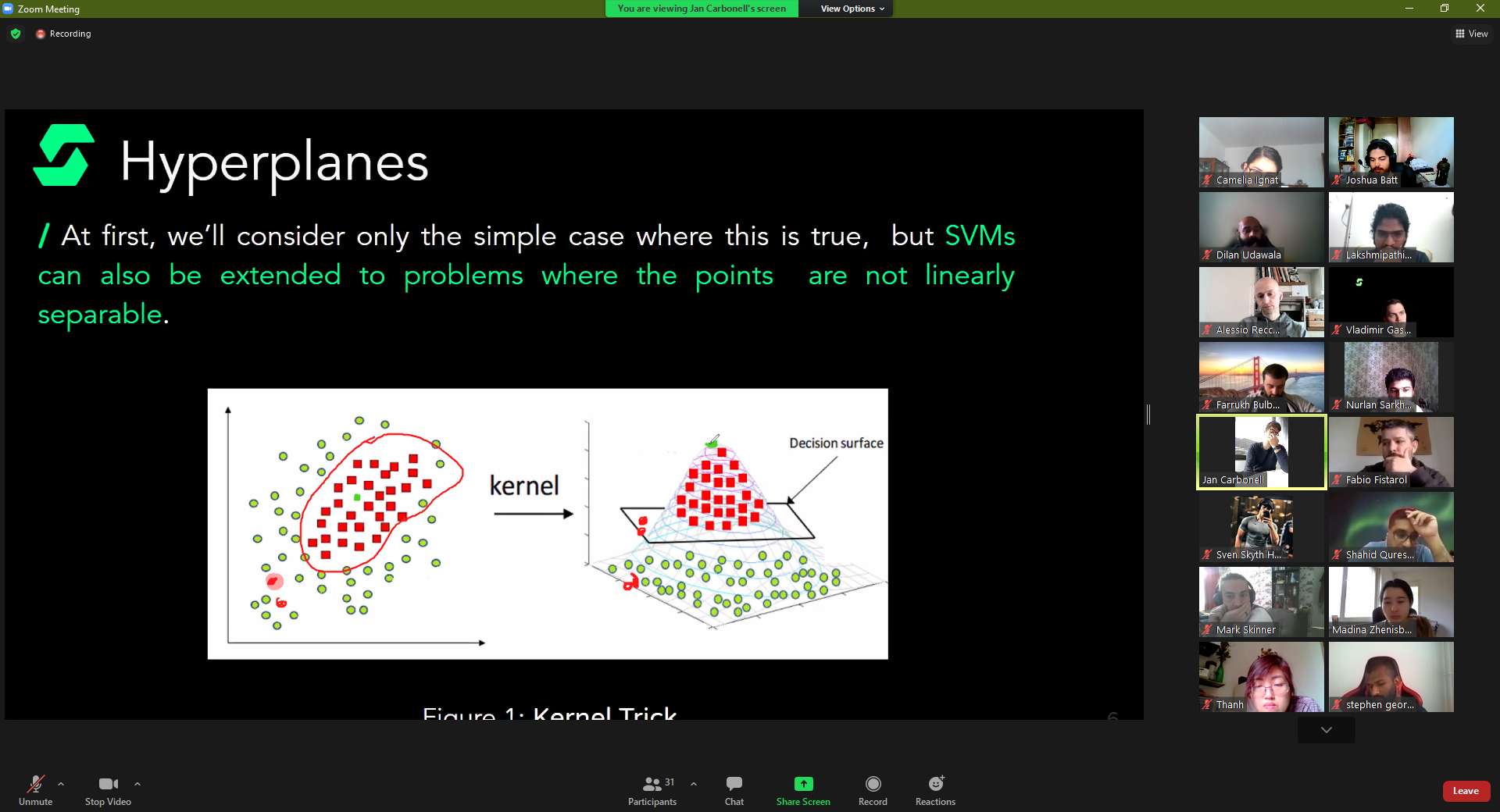
**M3 D9**

**SVM:**

* Stands for **S**upport **V**ector **M**achine (SVM)
* Tries to create the boundary between two data clusters.
* Only cares about the data points closest to boundary (the data points that are closest to the other cluster on the edge of their clusters)
  + 
* Good support vectors provide the widest possible margins.
  + 
* If you go for the smallest margin, you end up with boundary’s that don’t represent the clusters accurately
  + 
* Algorithm starts with smallest margins, then expands till it makes a mistake, then corrects till it has the largest possible margins between the two clusters.
* Boundary line is called hyperplane because it can act in any dimension.
  + 
* How to split data clusters that are in the middle of another cluster:
  + Kernel Trick
  + Make the data 3D temporarily, put a hyperplane in between the elevated bit and the ground.
  + Elevated part will be the data cluster that’s in the middle of the other cluster.
  + 
  + Only lifts the cluster, not the colour of dots:
    - 
  + **Projects the data to a higher-level dimension(s)**, it **does not transform the data**.
    - **If it transformed the data, this would then fall into the curse of dimensionality**.
* asd
* **EXTRA NOTE: KMeans does not consider boundaries, it just looks at the cluster and tries to find the middle of the cluster by calculating the means of each data point.**

**Data Preprocessing:**

* Three step process
  + Data Cleaning
  + Data Transformation
  + Data Reduction
* Data cleaning:
  + Keep eye out for placeholders (instead of NaN or 0 they put 999 or ?), if you don’t spot them, they won’t be delt with.
    - Must read dataset description carefully, if non available then check dataset manually
  + Only removes entries (whole rows/columns) that contain missing values if you have a large dataset and multiple (many!) values are missing
  + OR, fill the missing values. Either fill in manually, by attribute mean or the most probable value.
  + GO BACK AND READ THE SLIDES