

## Wrap up – Day I

Janis Keuper

## Basic Types of Machine Learning Algorithms

**Supervised Learning**

**Unsupervised Learning**

**Reinforcement Learning**

- Labeled data
- Direct and quantitative evaluation
- Learn model from „ground truth“ examples
- Predict unseen examples

## Basic Types of Machine Learning Algorithms

**Supervised Learning**

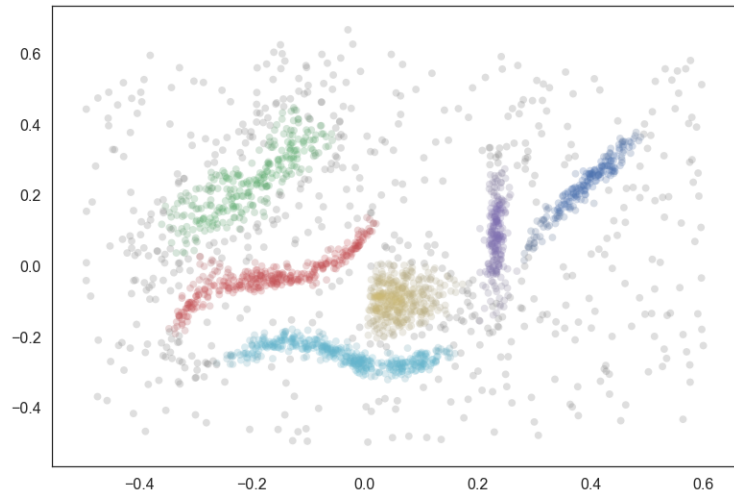
**Unsupervised Learning**

**Reinforcement Learning**

- NO Labeled data
- NO Direct and quantitative evaluation
- Explore structure of data

## Introduction

**Cluster analysis** or **clustering** is the task of **grouping a set of objects** in such a way that objects in the same group (called a cluster) are **more similar** (in some sense) to each other than to those in other groups (clusters). [Wikipedia]



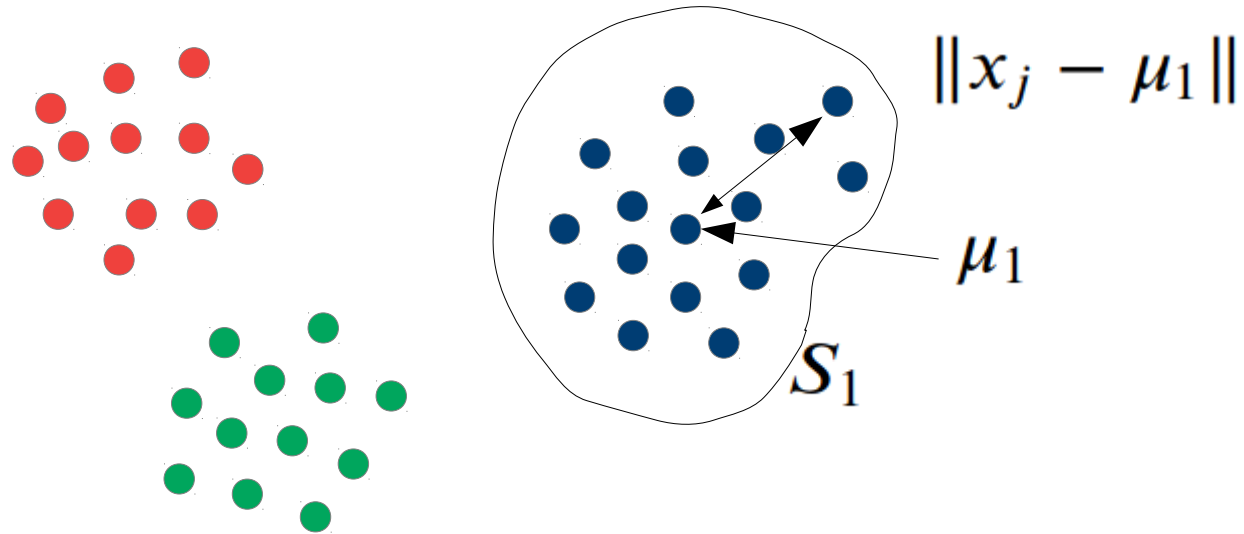
Example 2d data set

# Clustering Algorithms: K-Means

Intuition:

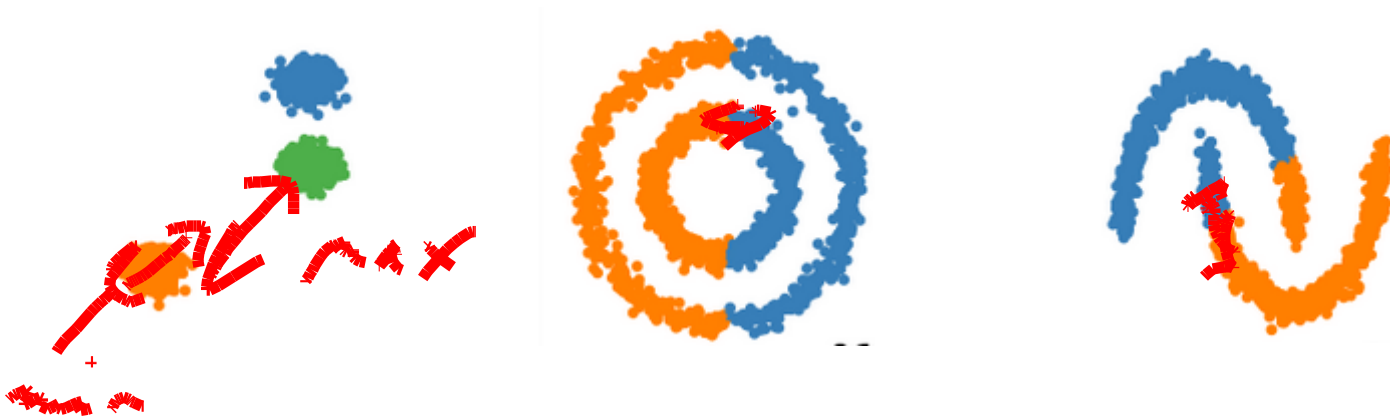
$$\arg \min[S] \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|$$

Clustering  
for  $k=3$



# Clustering Algorithms: K-Means

Evaluation:

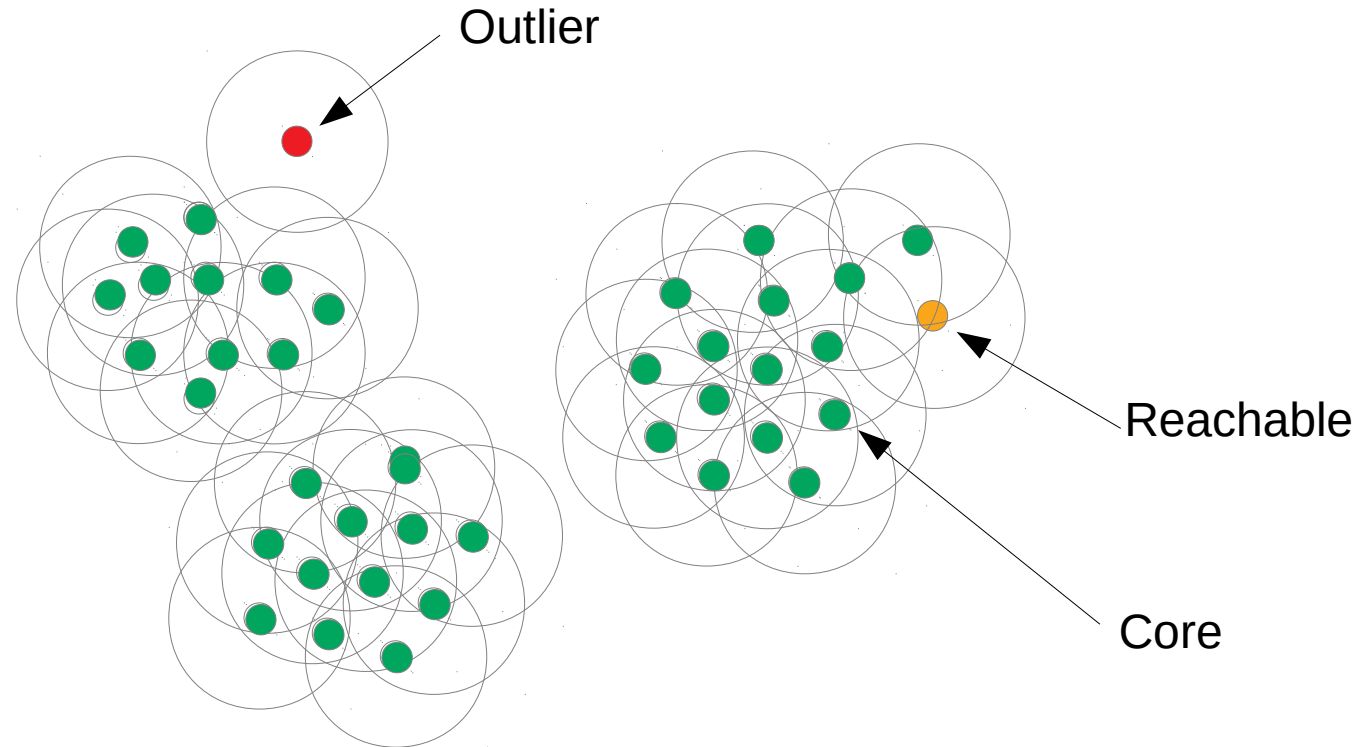
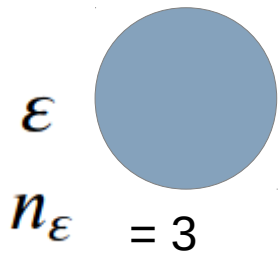


**More practical examples in the Lab session.... Now!**

# Clustering Algorithms: DBSCAN

Core – Reachable - Outlier:

Init:



# Clustering Algorithms: DBSCAN

**Evaluation:**



**More practical examples in the Lab session.... Now!**



## How to evaluate clustering:

- Visually → use dimension reduction techniques to visualize 2d or 3d

## How to evaluate clustering:

- Visually → use dimension reduction techniques to visualize 2d or 3d
- Quantitative quality measures (what is a good cluster?)

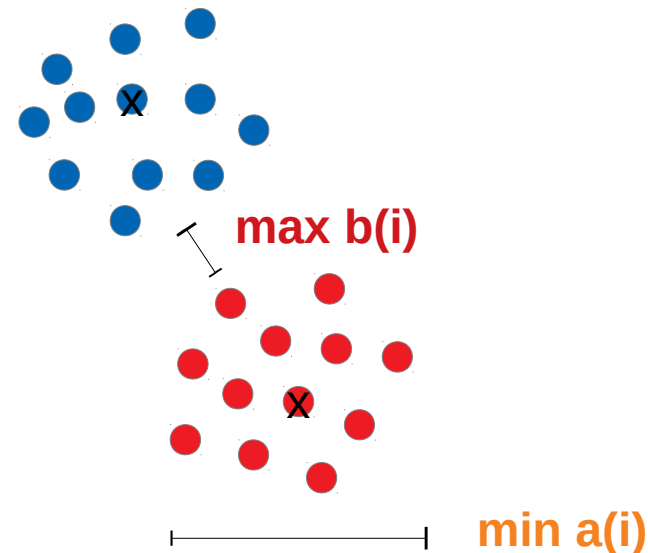
- **Low intra cluster variance**

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

- **High extra cluster variance**

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

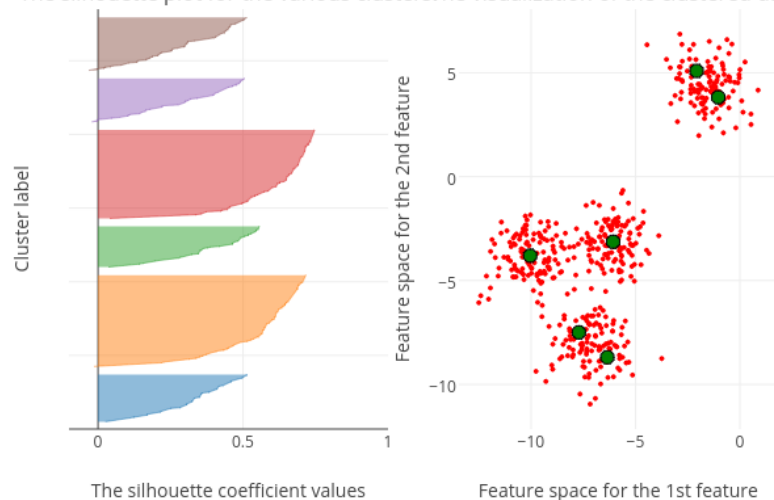
For each data point  $i \in C_i$  (data point  $i$  in the cluster  $C_i$ )



## Silhouette Diagrams: finding the best number of clusters

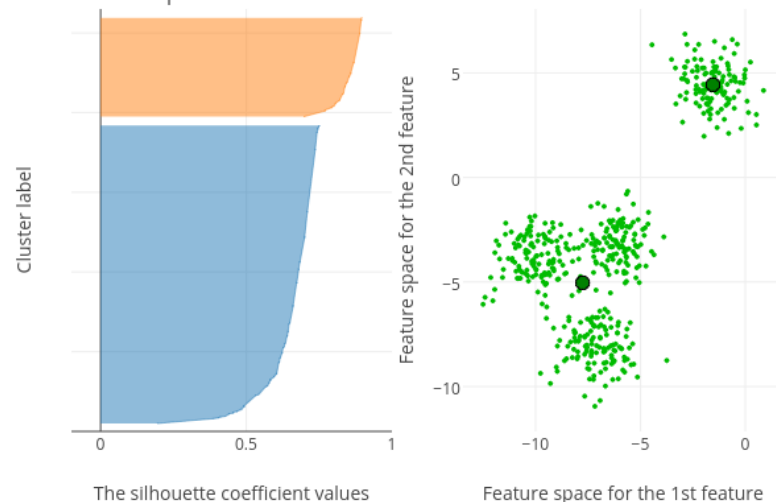
Silhouette analysis for KMeans clustering on sample data with  $n\_clusters = 6$

The silhouette plot for the various clustersThe visualization of the clustered data.

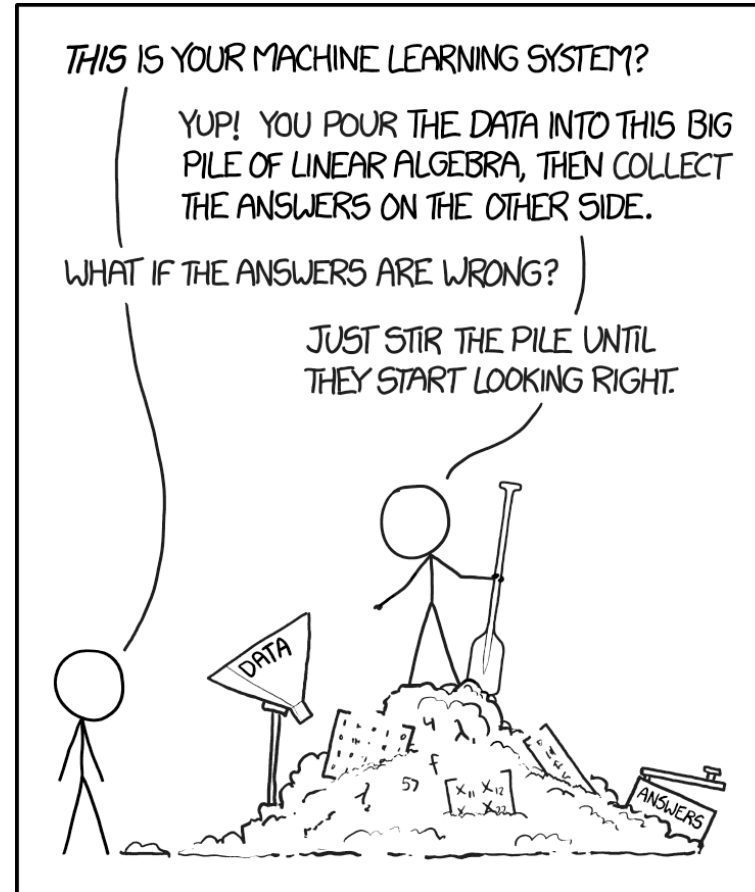


Silhouette analysis for KMeans clustering on sample data with  $n\_clusters = 2$

The silhouette plot for the various clustersThe visualization of the clustered data.

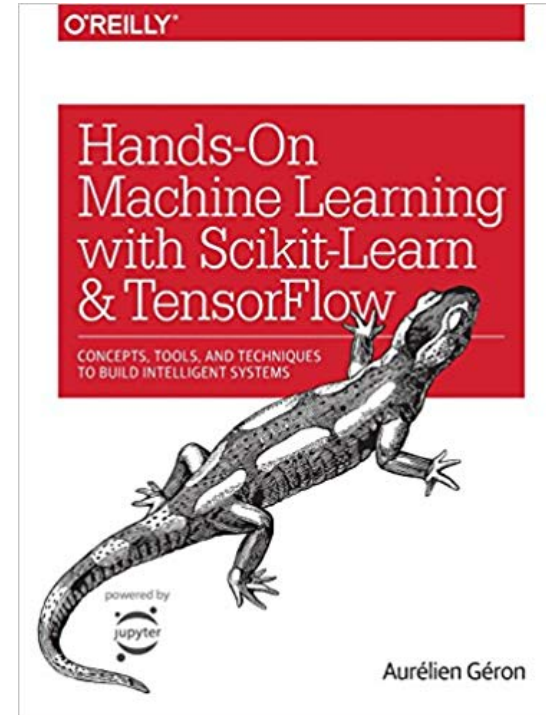
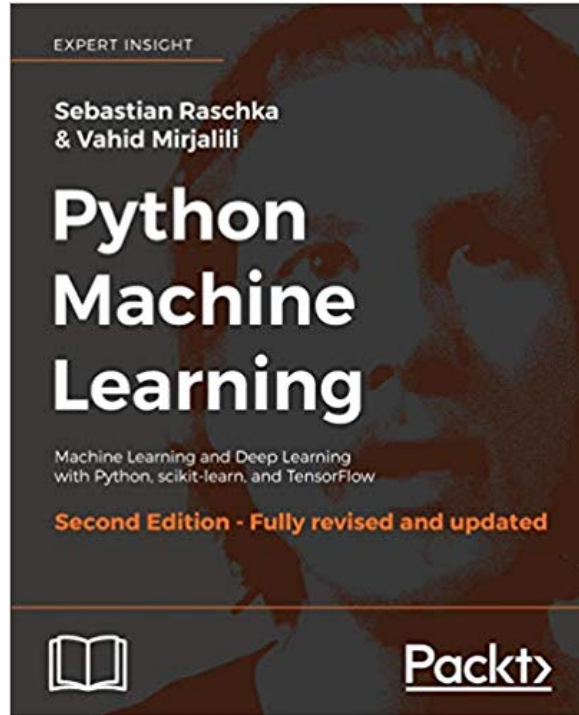


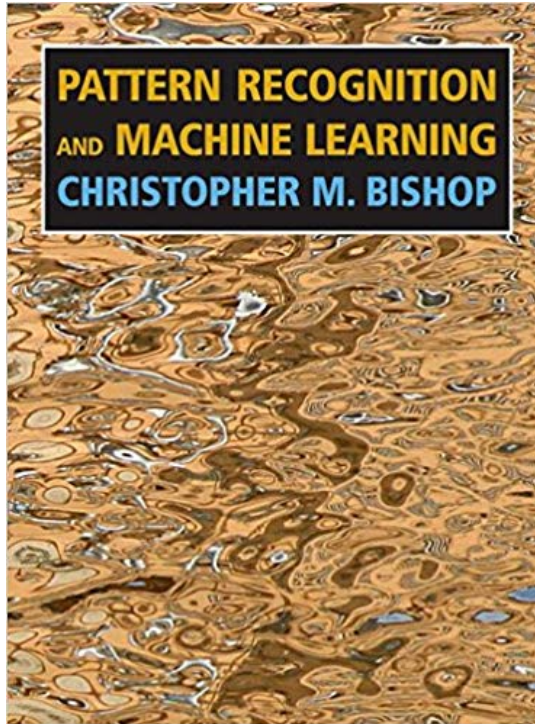
[plots: <https://plot.ly/scikit-learn/plot-kmeans-silhouette-analysis/>]



<https://xkcd.com/1838/>

„Hands on“ books:

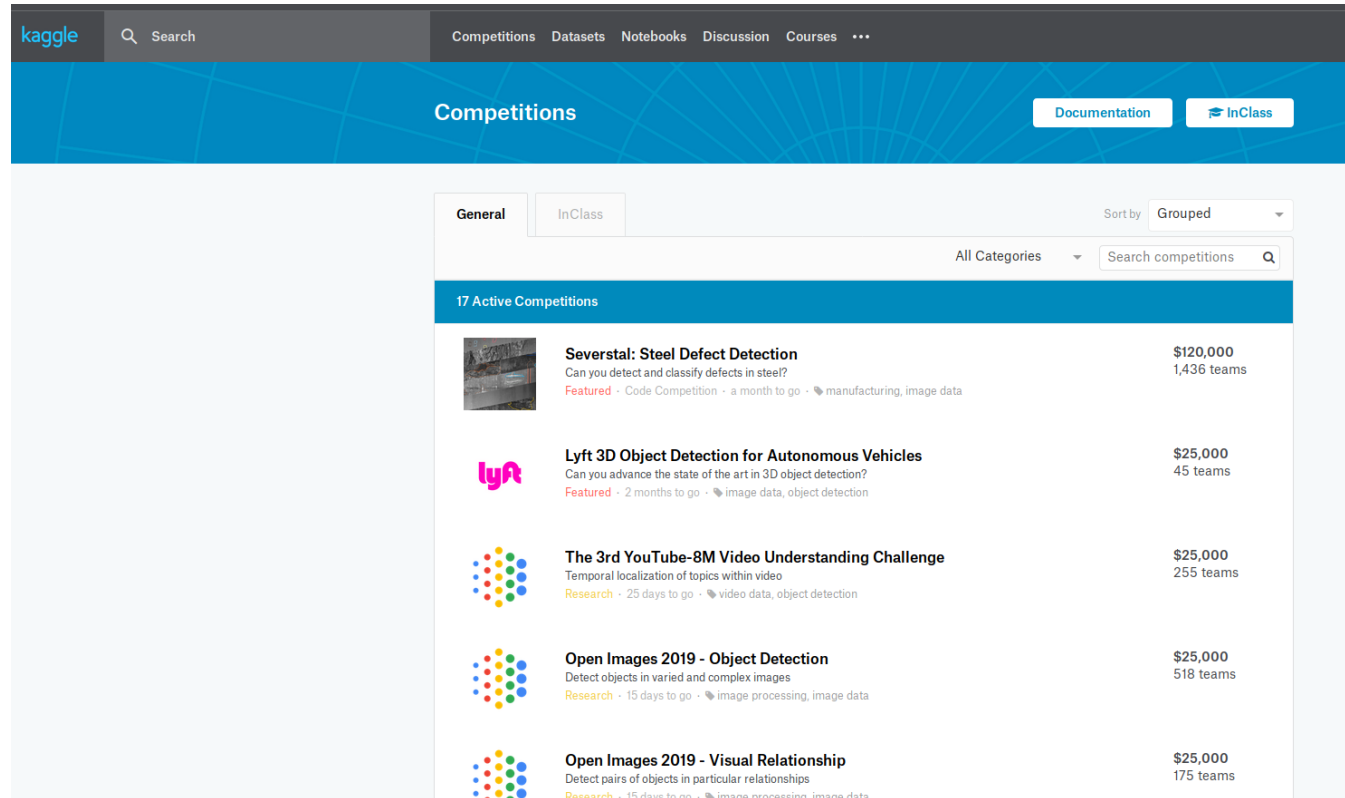




„complete“ theory

Learning by doing:

www.kaggle.com



The screenshot shows the Kaggle website's 'Competitions' section. The header includes the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Notebooks, Discussion, and Courses. The main content area is titled 'Competitions' and features a 'Documentation' button and an 'InClass' button. Below this, there are tabs for 'General' and 'InClass', and a 'Sort by' dropdown set to 'Grouped'. A search bar for competitions is also present. The main list displays '17 Active Competitions'. The first five are:

Competition	Prize	Teams
<b>Severstal: Steel Defect Detection</b> Can you detect and classify defects in steel? <i>Featured</i> · Code Competition · a month to go · manufacturing, image data	\$120,000	1,436 teams
<b>Lyft 3D Object Detection for Autonomous Vehicles</b> Can you advance the state of the art in 3D object detection? <i>Featured</i> · 2 months to go · image data, object detection	\$25,000	45 teams
<b>The 3rd YouTube-8M Video Understanding Challenge</b> Temporal localization of topics within video <i>Research</i> · 25 days to go · video data, object detection	\$25,000	255 teams
<b>Open Images 2019 - Object Detection</b> Detect objects in varied and complex images <i>Research</i> · 15 days to go · image processing, image data	\$25,000	518 teams
<b>Open Images 2019 - Visual Relationship</b> Detect pairs of objects in particular relationships <i>Research</i> · 15 days to go · image processing, image data	\$25,000	175 teams