Introduction to artificial intelligence and machine learning

Hands-on computer lab with Orange Data Mining

## Prerequisites

This lab will make use of the Orange data mining software application. It is a free and open source software that you can download and install from <https://orangedatamining.com/download/>. Installation guides and tutorials to get started are available at <https://orangedatamining.com/getting-started/>.

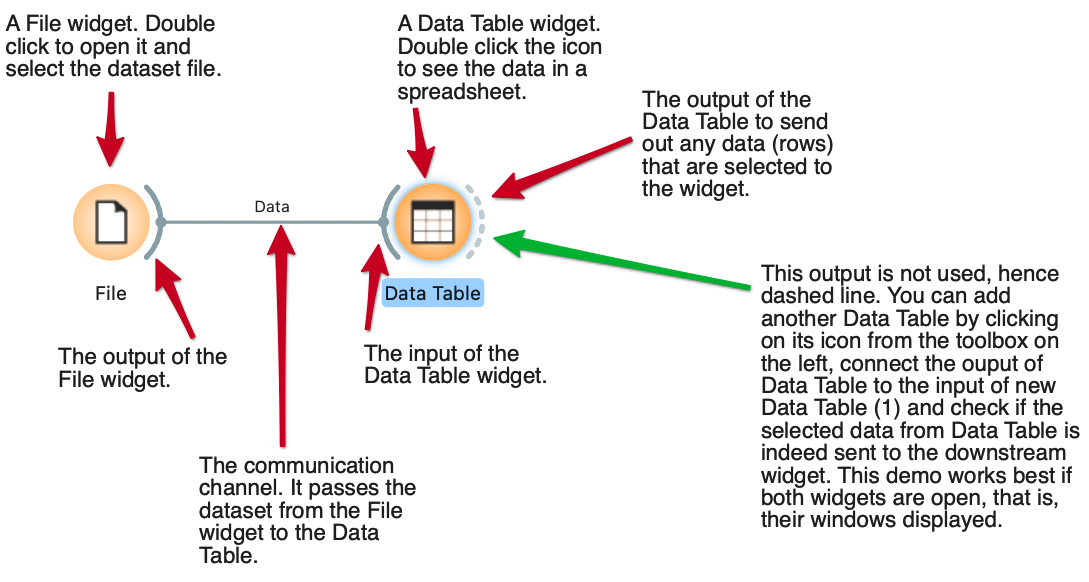
*Note: If you’re using Linux, remember to install in a Virtual Environment such as Conda, venv, pipenv etc. You will need to install addons through orange during this exercise - which requires elevated permissions.*

## 

## Exercise 1: Getting started with

**1.1 Widgets**

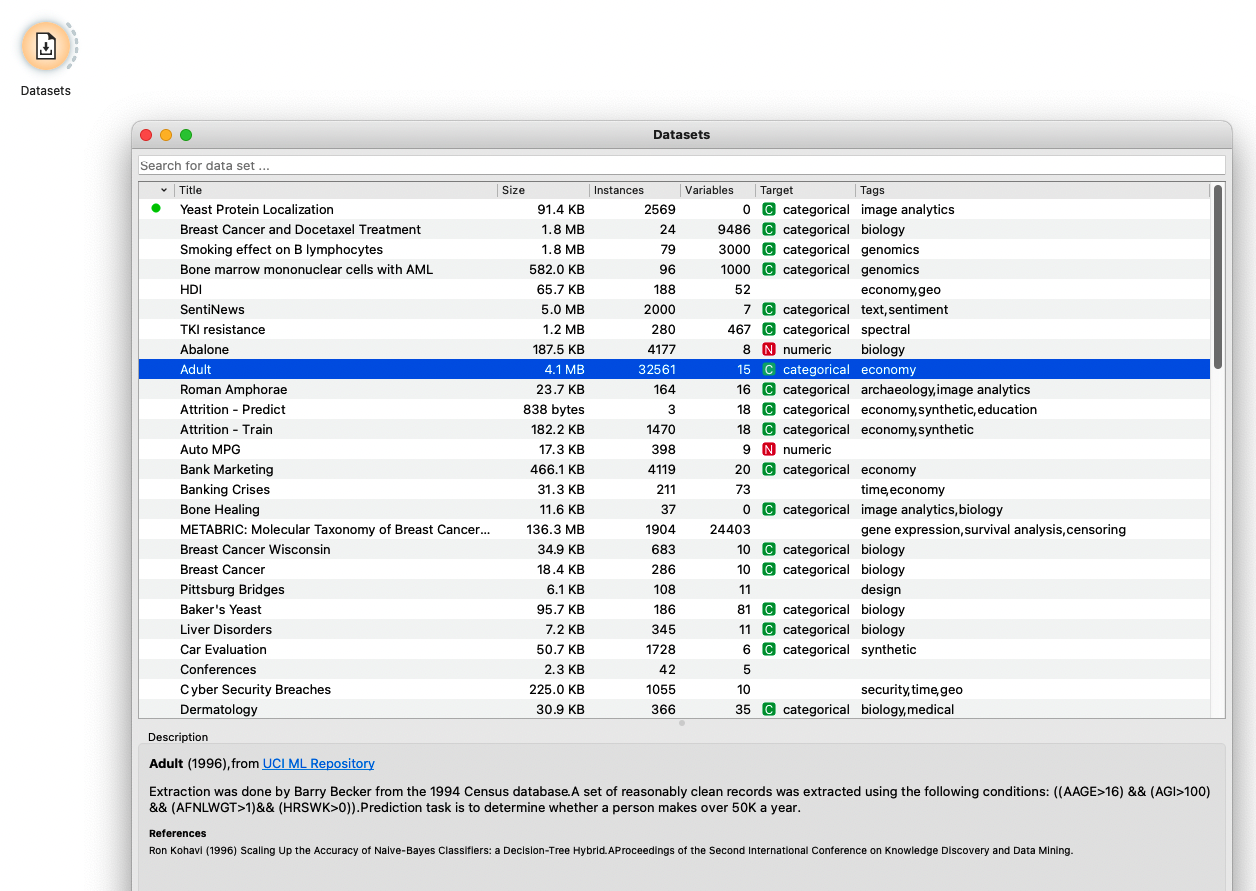
The visual components in Orange are called Widgets, and each one carries out some form of operation, for example loading data, executing an algorithm, or creating a plot. Widgets are connected through arrows referred to as Communication channels. Orange contains a lot of widgets, you can see a list of widgets in the Widget Catalog <https://orangedatamining.com/widget-catalog/>.



***Figure 1:*** *Basic operations in Orange.*

**1.2 Loading data**

Orange comes with a set of pre-made datasets. Click the **Datasets** widget in the left toolbar to add it to the workbench, then double-click it to bring up a list of datasets that can easily be imported and accessed.

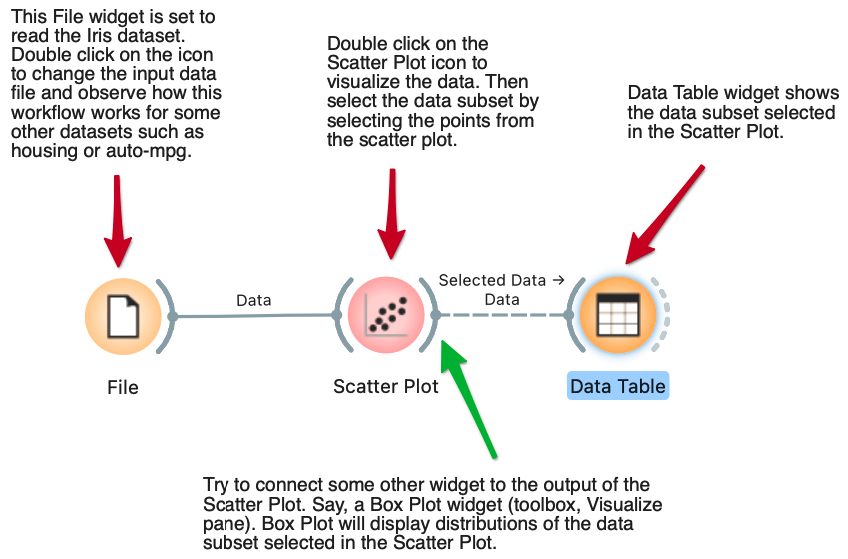


***Figure 2:*** *List of datasets shipped with Orange.*

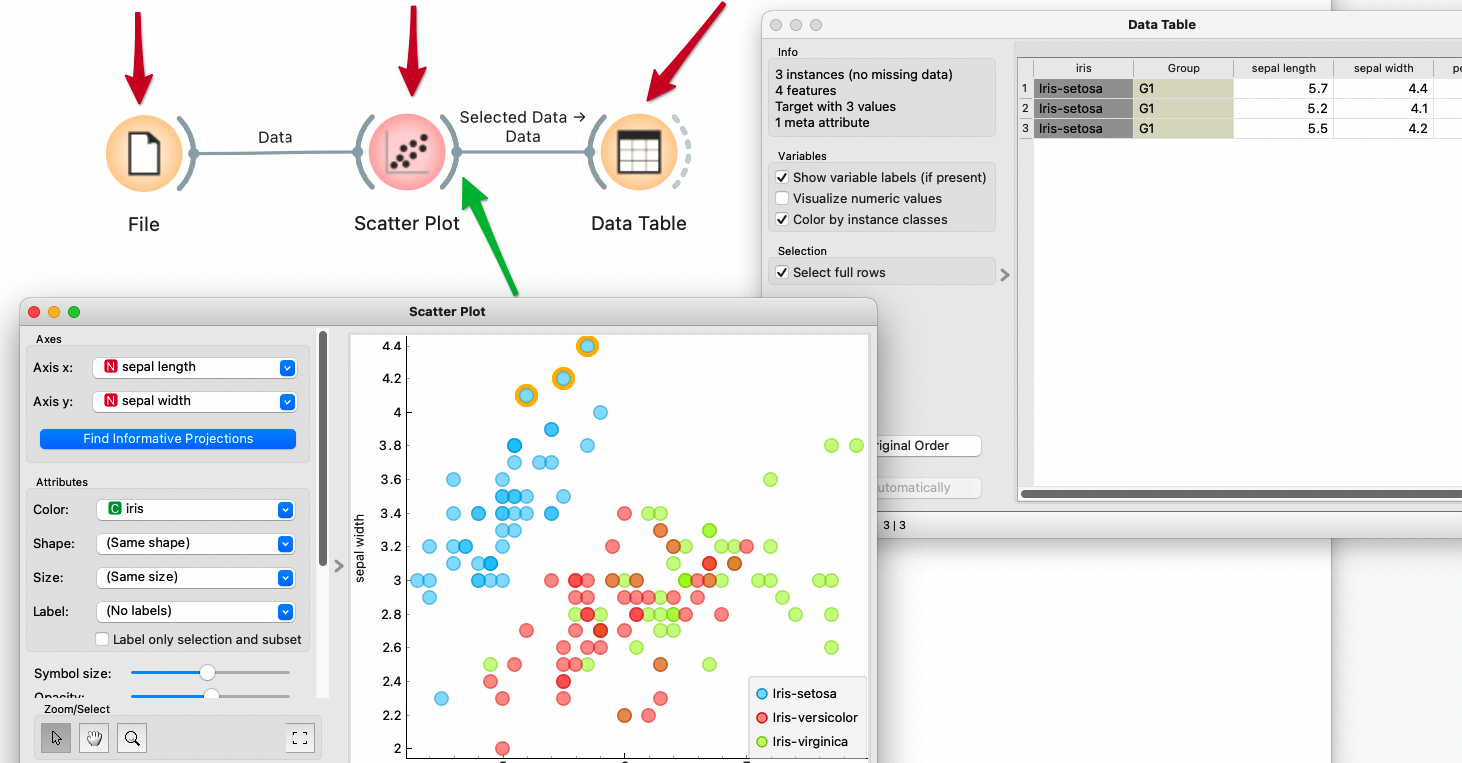
A very common task is otherwise to import data from a file. Use the **File** widget for standard spreadsheet files (e.g. Excel) or use the **CSV File Import** widget for comma-separated files (CSV) if you need to customize the separator between columns.

**1.3 Basic visualizations**

After loading data you can easily visualize the content in a plot or a data table. Create the workflow below and test the **Scatter Plot** and the **Data Table** widgets.



***Figure 3:*** *Example workflow for loading the Iris dataset, visualize the data in a* ***Scatter Plot****, and inspect the data in a* ***Data Table****.*



***Figure 4:*** *A scatter plot and a data table. Note that since the communication channel is set to “Selected Data” then the data table will only show the selected objects in the scatter plot.*

## Exercise 2: Unsupervised learning - Analysis of a phenotypic drug screen

**Background:**

The dataset we’ll use in this exercise originates from a drug repurposing screen using a modified version of the Cell Painting protocol [[1]](https://paperpile.com/c/XQXGjb/DBLWH) where VERO E6 cells are first infected by the SARS-CoV-2 virus and then exposed to existing drugs. After 24 hours exposure, cells are subjected to multiplexed fluorescence staining using the Cell Painting protocol [[2]](https://paperpile.com/c/XQXGjb/uKnf6) and imaged using automated high-content imaging. The experiments are carried out in 384-well plates, and 9 sites are imaged in each well in 5 channels capturing morphological changes to different compartments (DNA, nucleoli and cytoplasmic RNA, Actin filaments, ER, plasma membrane and Golgi), producing ca 15,000 images per plate. In this screen, an antibody was added to stain viral proteins allowing to measure the rate of infection on a single-cell basis [[1]](https://paperpile.com/c/XQXGjb/DBLWH).

The objective of the screen was to study changes in cell morphology and identify existing drugs that are able to reverse the disease phenotype, driving cells from infected state to non-infected state. The SPECS drug repurposing library of 5463 established drugs was screened, which together with DMSO (negative control) Remdesivir (reference compound) and uninfected cells made up a screen consisting of 32 plates, and resulted in approximately 500,000 images (4 TB storage). CellProfiler software [[3]](https://paperpile.com/c/XQXGjb/tKeMF) was used to calculate image-based features on a single-cell basis, and median values per compound were calculated. The resulting dataset has 6065 rows and 1881 features. The features are measures of cell morphology from the images of the different cell compartments (e.g. size of nucleus, texture of golgi, clustering of mitochondria etc). In this exercise we will use the same dataset but with a subset of the features to reduce the size and calculation times, but the signal from the original dataset is preserved.

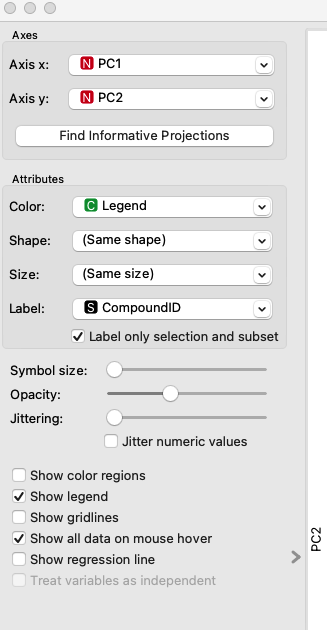
**Objective:** Identify drugs that drive cells from infected to uninfected state. Compare with the known antiviral Remdesivir.

**Outline:**

1. Download the zip-file lab1-orange from this google drive:

<https://drive.google.com/drive/folders/1C8tB0mHnxAnAaO-aWR6xebCOYT1ddbVo?usp=sharing>

1. Load the data file **COVID-CP-screen-edu.csv**
   1. Use the **CSV File Import** widget. Make sure to set the right delimiter.
2. Perform a PCA to reduce to 2 dimensions.
   1. How much of the variation in the data is explained?
3. Perform a scatter plot and inspect how the treated cells cluster. Each object (circle) in the plot corresponds to a treatment of cells.
   1. Does Remdesivir drive cells from infected to uninfected cell state?
   2. Hint: Ensure the Scatter plot has principal components 1 and 2 (PC1 and PC2) on the X and Y axis, and that you color by Legend (see figure below).



1. Locate the top 5 screened drugs that drive cells in the direction from infected towards uninfected cell state

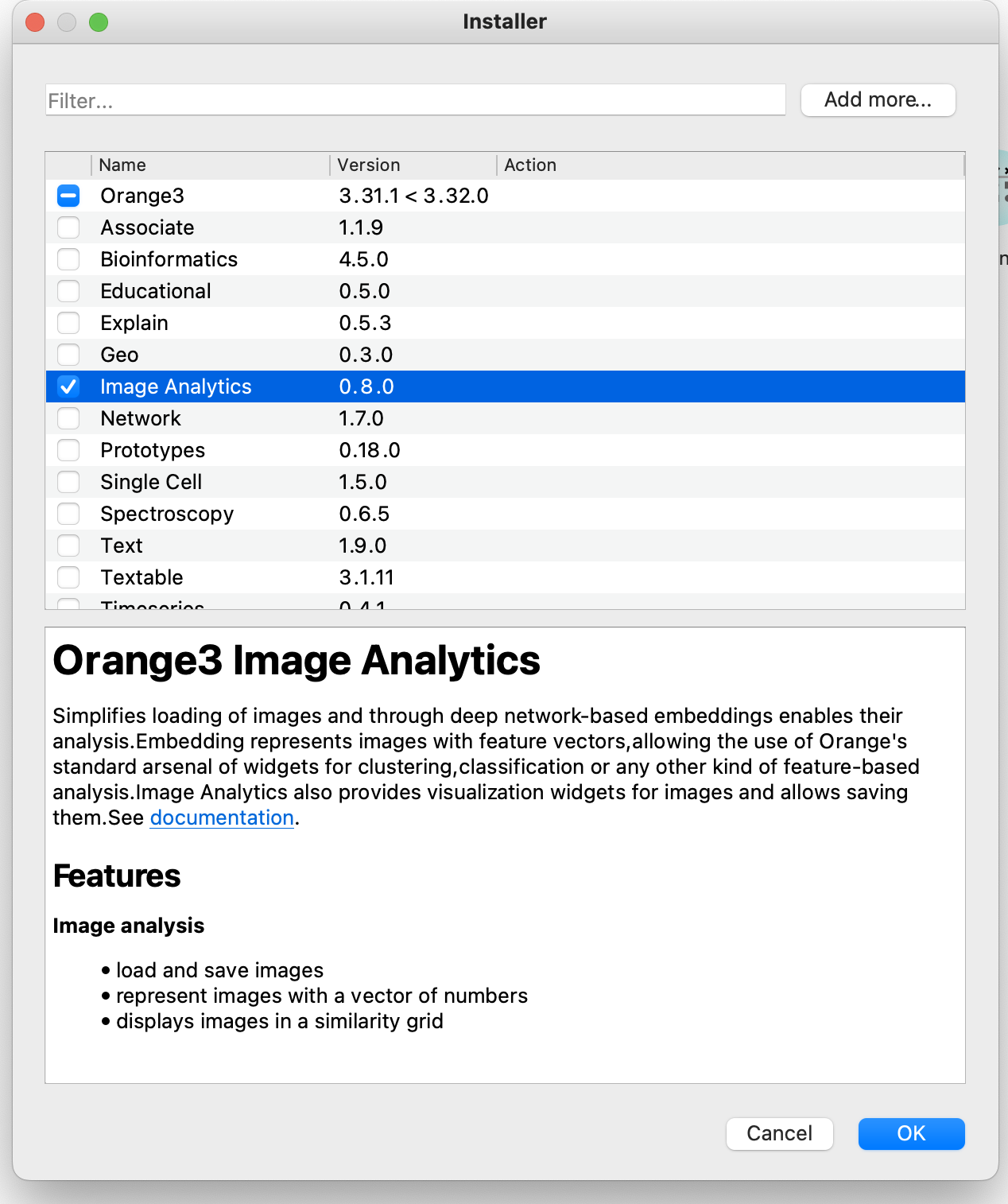
## Exercise 3: Supervised learning - Classifying yeast images into 5 categories

In this exercise we will train ML models to distinguish between images of yeast cells, more specifically where certain organelles are stained. We will use Deep Learning or more specifically Convolutional Neural Networks (CNNs) that are pre-trained on [ImageNet](https://image-net.org/) [[6]](https://paperpile.com/c/XQXGjb/HeRzk) for image embedding. The outcome is a vector representation of images, and we can then train ML models on this data.

**Objective**: Train a machine learning model to classify yeast images. Experiment with different representations and modeling methods.

Outline:

1. Install the Image Analytics add-on from the menu Options -> Add-ons…



1. Download and unzip the datafile yest-cells.zip from the google drive: <https://tinyurl.com/isr22ht>
2. From the Image Analytics section in the left table, add the Import Images widget and load the images by selecting the root folder yplplus that contains 5 subfolders. Each subfolder will be treated as a category. There will be 5 categories: CYTOPLASM, ENDOSOME, ER, MITOCHONDRIA, NUCLEUS.
3. Add an Image Viewer and inspect the images.
4. Add an Image Embedding widget. This widget contains pre-trained models (convolutional neural networks) on ImageNet that are generally good for creating image embeddings.
5. Inspect the output from the Image Embedding widget. What is the size of the embedding (feature vector)?
6. Train a kNN model on the data. What is the AUC? F1?
7. Add a Confusion Matrix and after that connect another Image Viewer to it. Select misclassified objects in the Confusion matrix and inspect them in the Image Viewer.
8. Test different embeddings and modeling methods (e.g. RF, SVM, Logistic Regression) and compare their F1 scores. You can read more about types of CNNs here: <https://medium.com/analytics-vidhya/types-of-convolutional-neural-networks-lenet-alexnet-vgg-16-net-resnet-and-inception-net-759e5f197580>
   1. What combination of embedding/modeling method gives the best F1 score?

## References

1. [Rietdijk J, Tampere M, Pettke A, Georgiev P, Lapins M, Warpman-Berglund U, et al. A phenomics approach for antiviral drug discovery. BMC Biol. 2021;19: 156.](http://paperpile.com/b/XQXGjb/DBLWH)

2. [Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016;11: 1757–1774.](http://paperpile.com/b/XQXGjb/uKnf6)

3. [McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol. 2018;16: e2005970.](http://paperpile.com/b/XQXGjb/tKeMF)

4. [Michalski, Mozetic, Hong, Lavrac. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. Proc AAAI. 1986. Available:](http://paperpile.com/b/XQXGjb/qcCt) <https://www.aaai.org/Papers/AAAI/1986/AAAI86-171.pdf>

5. [Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.](http://paperpile.com/b/XQXGjb/K9o7H)

6. [Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 248–255.](http://paperpile.com/b/XQXGjb/HeRzk)