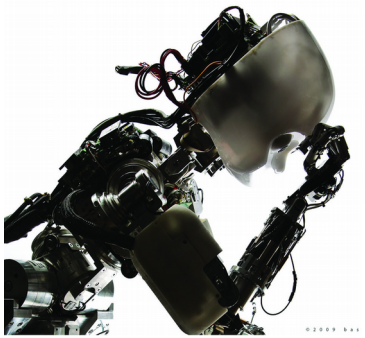
ADVANCED MACHINE LEARNING

Mini Project - Dataset

Lecture : Prof. Aude Billard (aude.billard@epfl.ch)

Teaching Assistants :
Lukas Huber, Bernardo Fichera, ,Thomas Pethick
lukas.huber@epfl.ch / bernardo.fichera@epfl.ch / thomas.pethick@epfl.ch



LASA
Learning Algorithms and Systems Laboratory

EPFL

Dataset Selection

There is no *wrong* dataset.

> The mini-project is very open. It allows for different analysis for different dataset.

i.e. a dataset which requires more preprocessing → this should be more important int the report

Clustering

> No label needed
> Class need to be clearly separable

Classfication

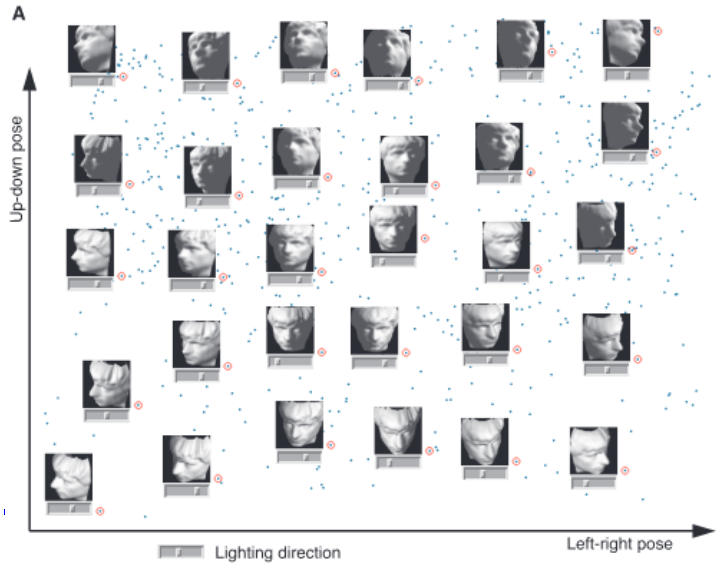
> Label needed!
> Data can be more overlying than clustering

Regression

> Continuous dataset
> Prediction inside

Manifold Learning

> Extracting the differences/distribution of the dataset
e.g. pixels > head pose

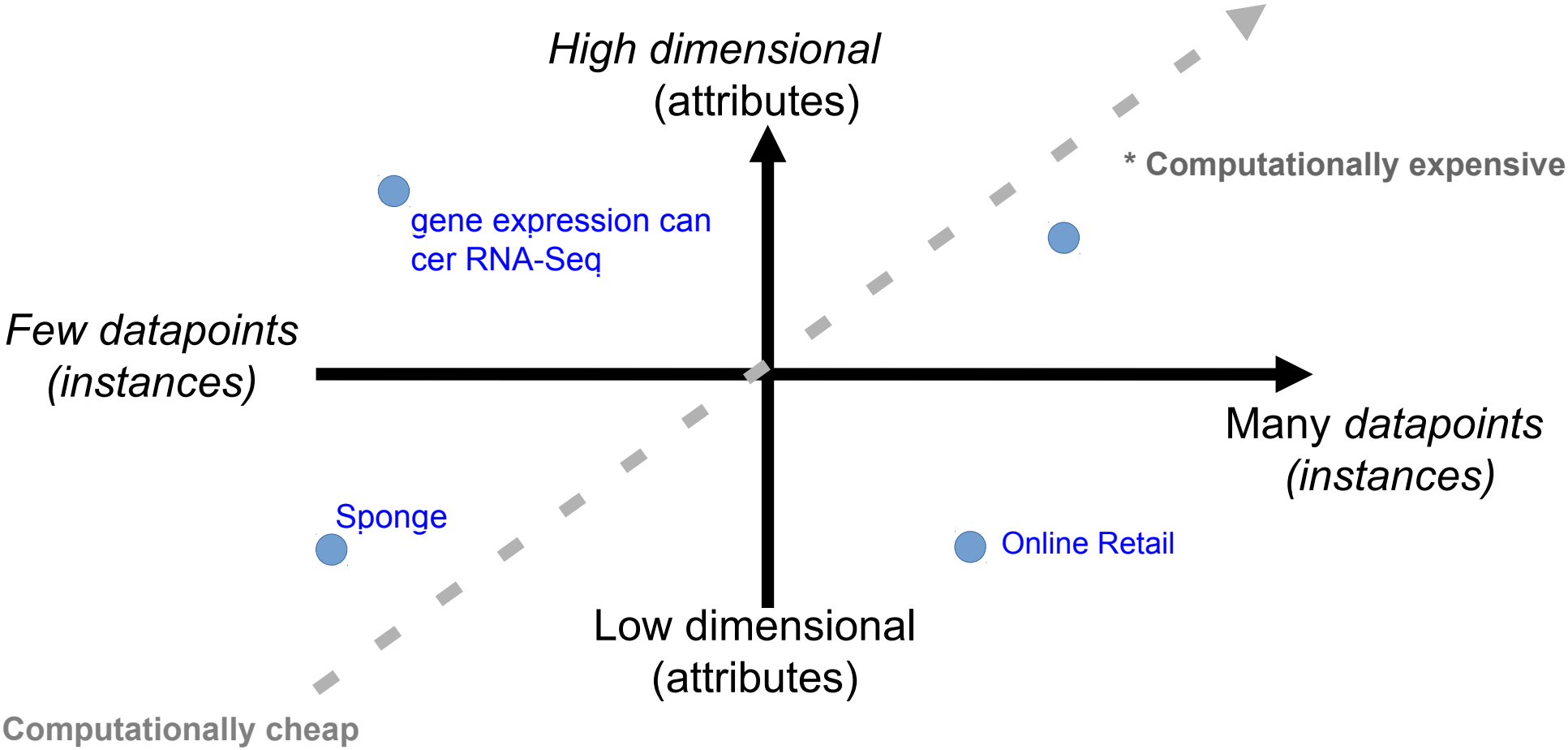


Example Datasets:
Handwritten Digits
http://yann.lecun.com/exdb/mnist/index.html

Facecrumbs
http://www2.ece.ohio-state.edu/~aleix/ARdatabase.

AR Face
https://kevin-keraudren.github.io/facecrumbs.html

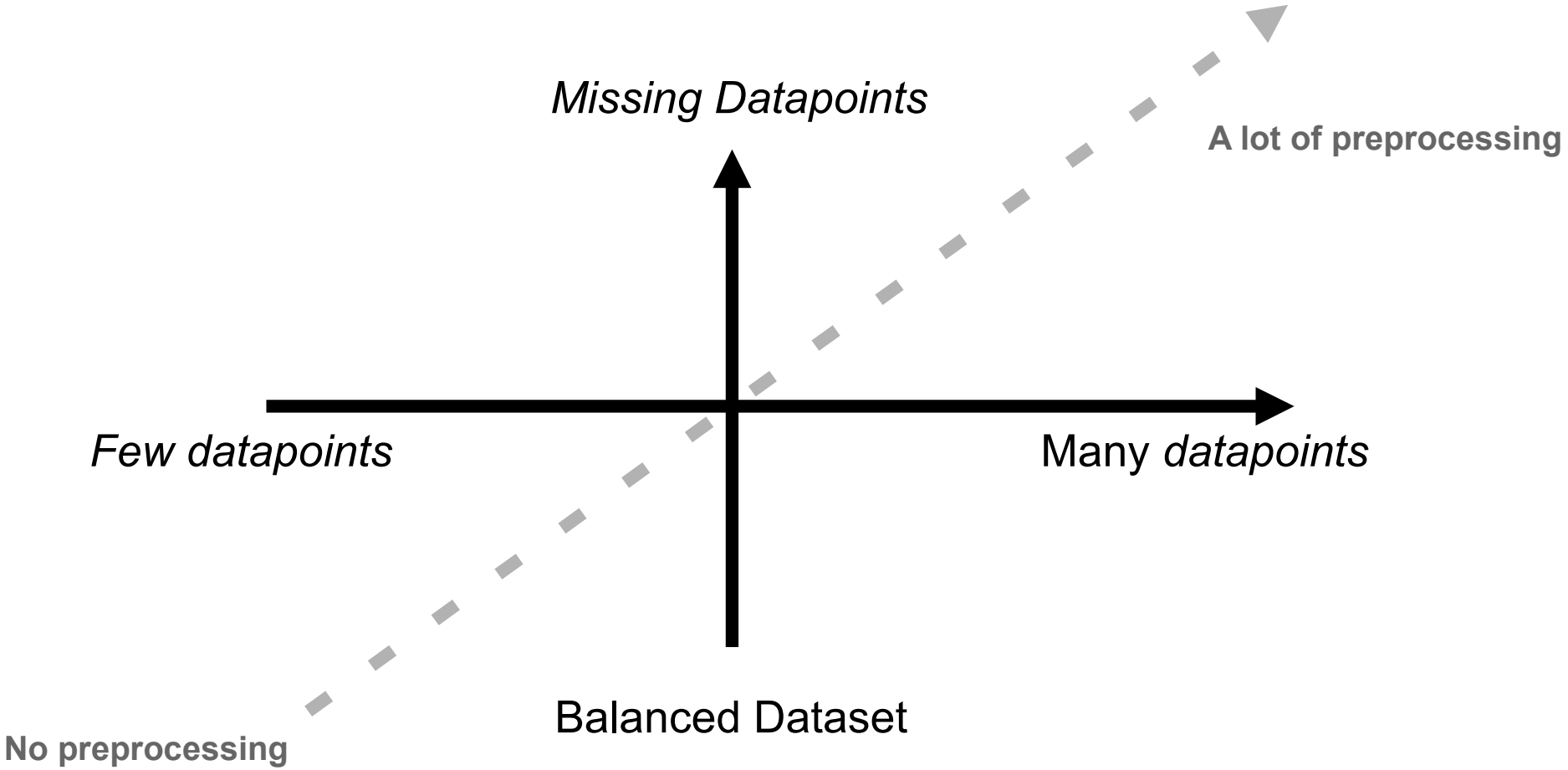
Dataset Selection



*Google Provides free CPU for research
https://colab.research.google.com

Source: https://archive.ics.uci.edu/ml/datasets.php

Dataset Selection



Visualize in a first step your data set, to see verify what your dataset looks like. (Change dataset if desired.)

Source: https://archive.ics.uci.edu/ml/datasets.php

Dataset Selection (UCI)

Computational Cost

Your algorithm

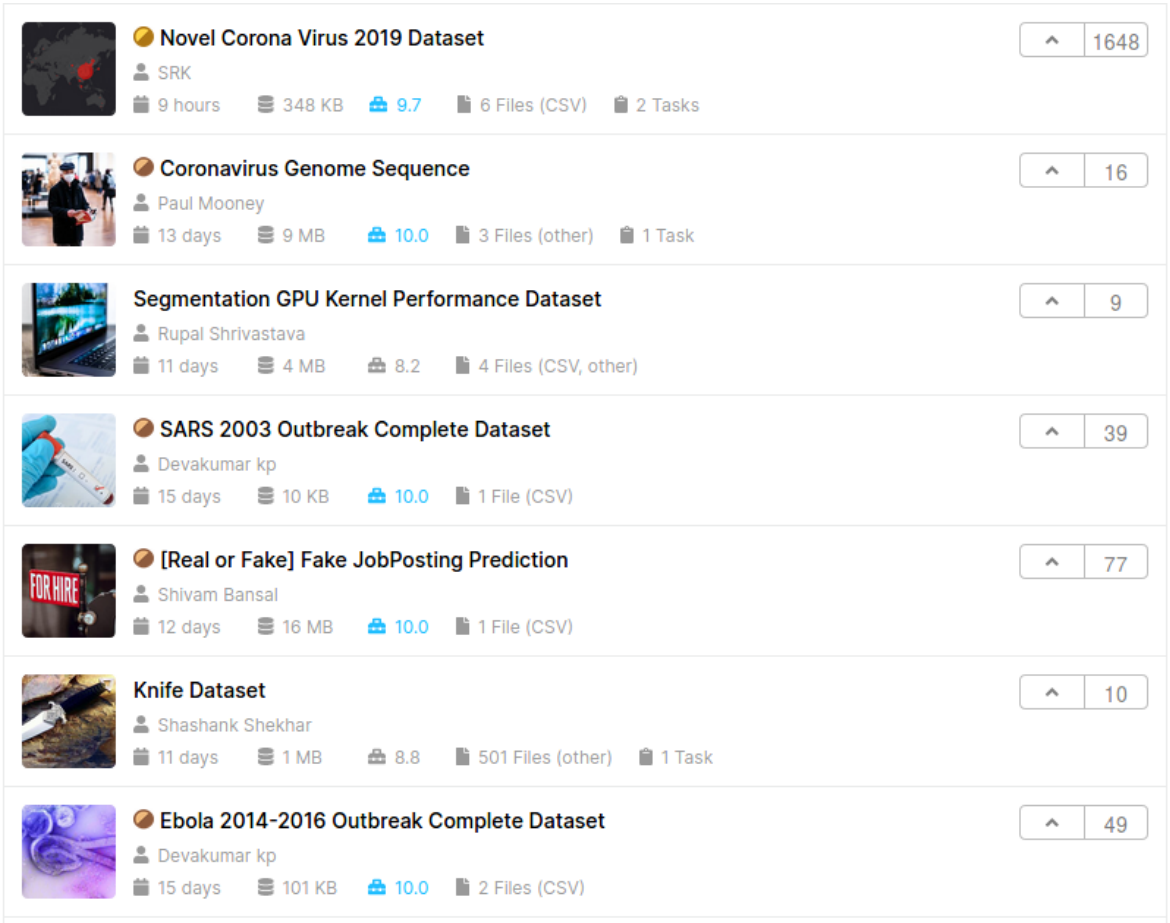
Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
3D Road Network (North Jutland, Denmark)	Sequential, Text	Regression, Clustering	Real	434874	4	2013
A study of Asian Religious and Biblical Texts	Multivariate, Text	Classification, Clustering	Integer	590	8265	2019
AAAI 2013 Accepted Papers	Multivariate	Clustering		150	5	2014
AAAI 2014 Accepted Papers	Multivariate	Clustering		399	6	2014
Absenteeism at work	Multivariate, Time-Series	Classification, Clustering	Integer, Real	740	21	2018

Preprocessing needed?
(see next week)

Source: https://archive.ics.uci.edu/ml/datasets.php

Dataset Selection (Kaggle)

More recent (interesting?) data → less support on choosing



Source: https://www.kaggle.com/datasets