

ADVANCED MACHINE LEARNING

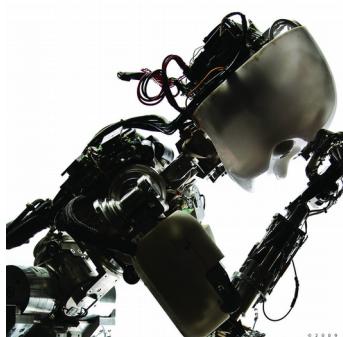
Mini Project - Dataset

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

Classification

- > 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

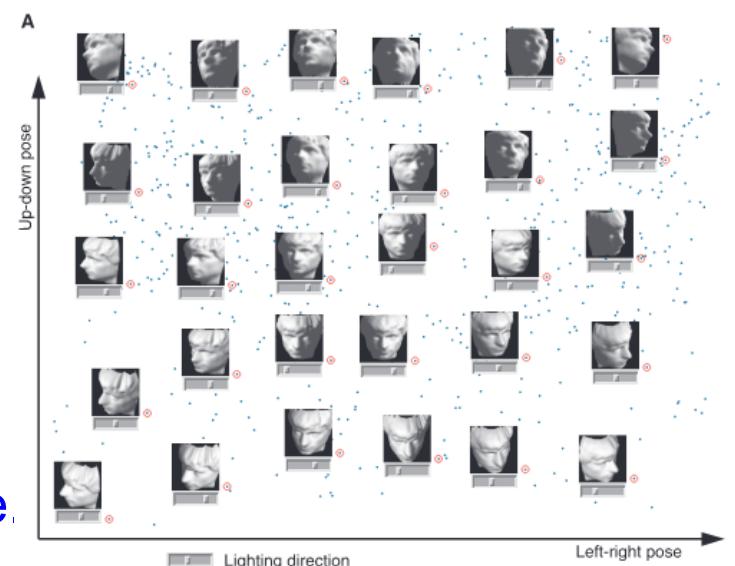
<http://yann.lecun.com/exdb/mnist/index.html>

Facecrumbs

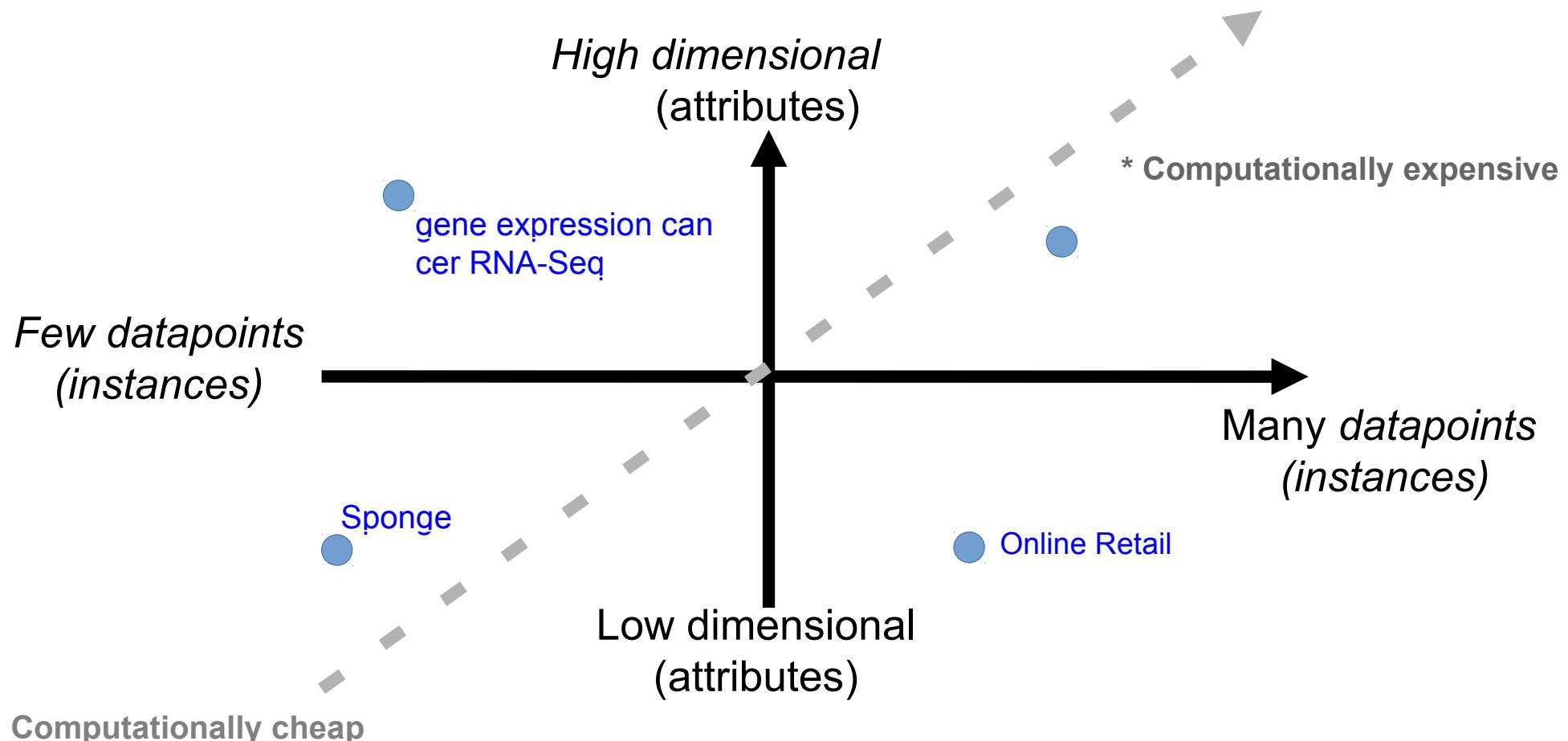
<http://www2.ece.ohio-state.edu/~aleix/ARdatabase>

AR Face

<https://kevin-keraudren.github.io/facecrumbs.html>



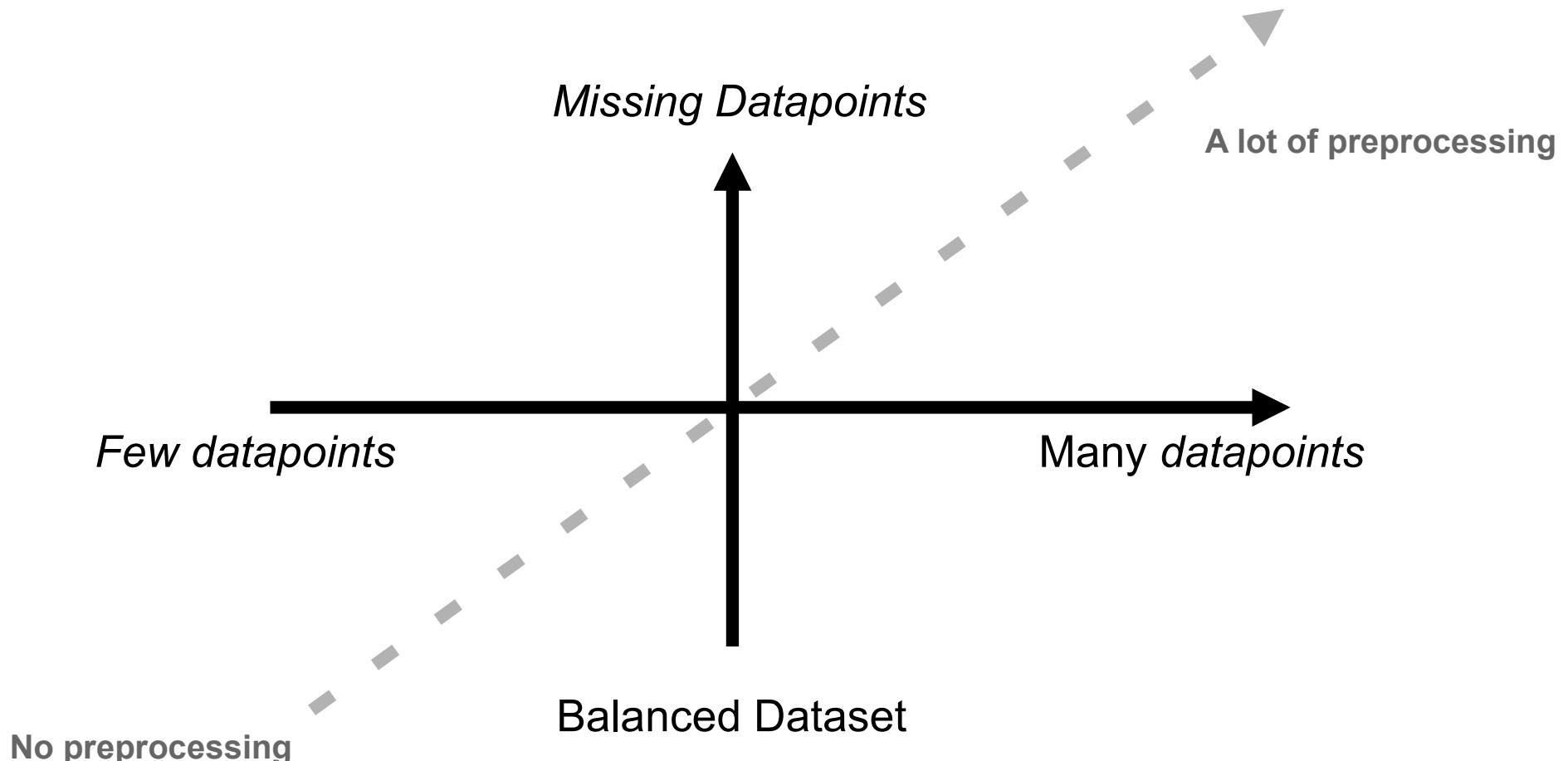
Dataset Selection



*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.)

Dataset Selection (UCI)

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

Your algorithm

Computational Cost

Preprocessing needed?
(see next week)

Dataset Selection (Kaggle)

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

	Novel Corona Virus 2019 Dataset by SRK 9 hours ago · 348 KB · 9.7 · 6 Files (CSV) · 2 Tasks	1648
	Coronavirus Genome Sequence by Paul Mooney 13 days ago · 9 MB · 10.0 · 3 Files (other) · 1 Task	16
	Segmentation GPU Kernel Performance Dataset by Rupal Shrivastava 11 days ago · 4 MB · 8.2 · 4 Files (CSV, other)	9
	SARS 2003 Outbreak Complete Dataset by Devakumar kp 15 days ago · 10 KB · 10.0 · 1 File (CSV)	39
	[Real or Fake] Fake JobPosting Prediction by Shivam Bansal 12 days ago · 16 MB · 10.0 · 1 File (CSV)	77
	Knife Dataset by Shashank Shekhar 11 days ago · 1 MB · 8.8 · 501 Files (other) · 1 Task	10
	Ebola 2014-2016 Outbreak Complete Dataset by Devakumar kp 15 days ago · 101 KB · 10.0 · 2 Files (CSV)	49

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