MACHINE LEARNING II

Course Themes

Practicalities & Infos

Aude Billard

Structure Discovery

Goal: discover (possibly low-dimensional) structure of data.

Method**:**

Use **nonlinear transformation** to determine (lower-dimensional or higher-dimensional) representations of the dataset (**feature space** or **latent space**); project or lift data in this new space

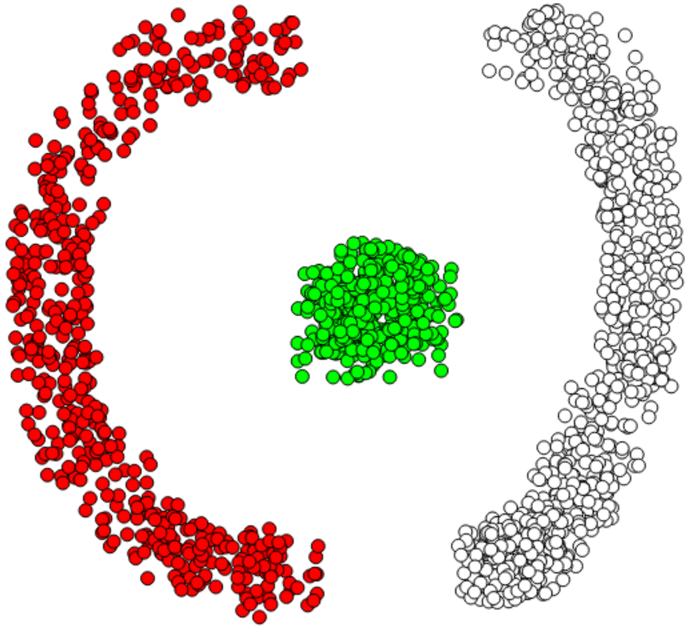
Determine **descriptors** of the new representation; study the data in this new representation

Principle:

The **new representations and descriptors reveal the underlying structure** of the data.

It can *allow one to visualize, denoise, and interpret the data*.

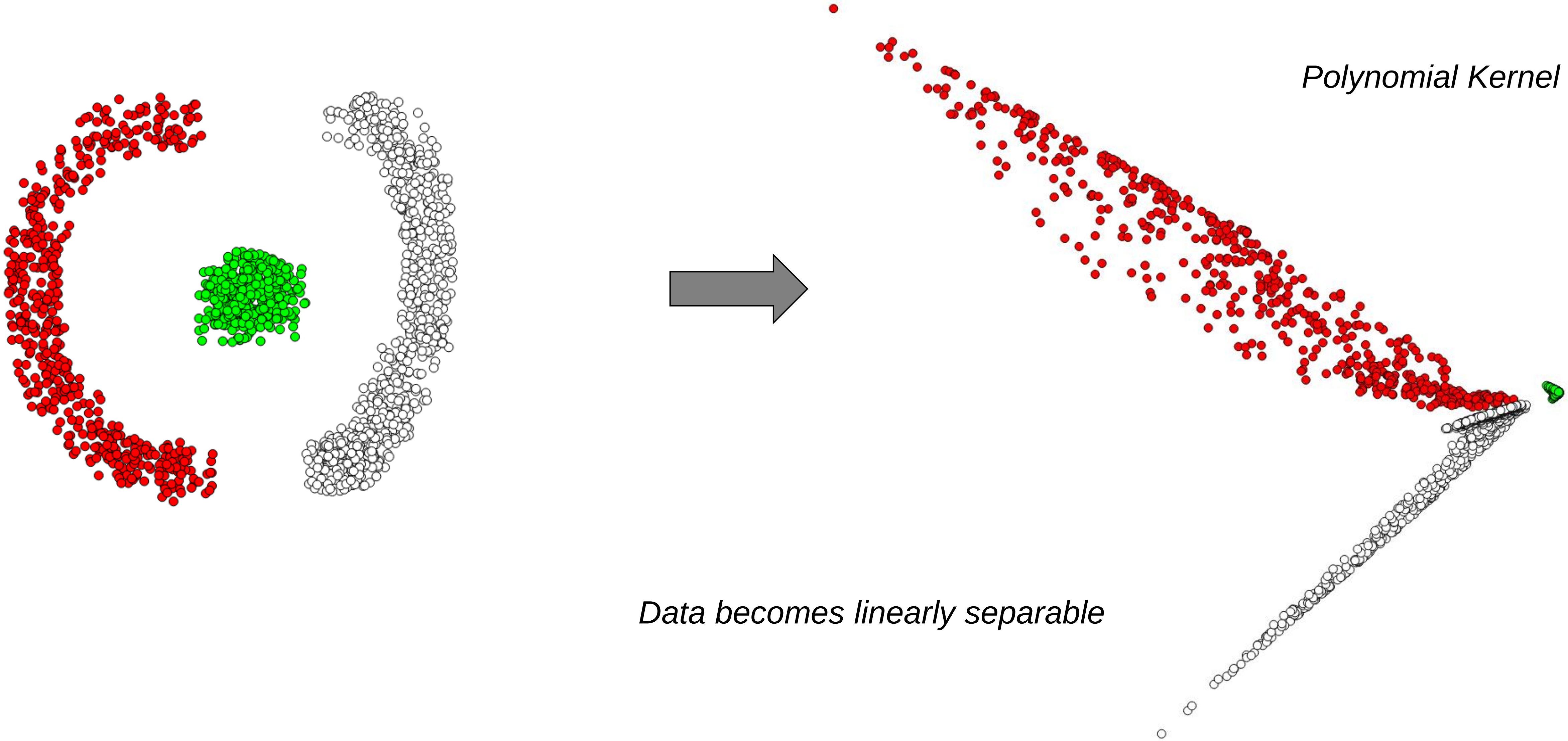
Kernel PCA



Polynomial Kernel

Data becomes linearly separable

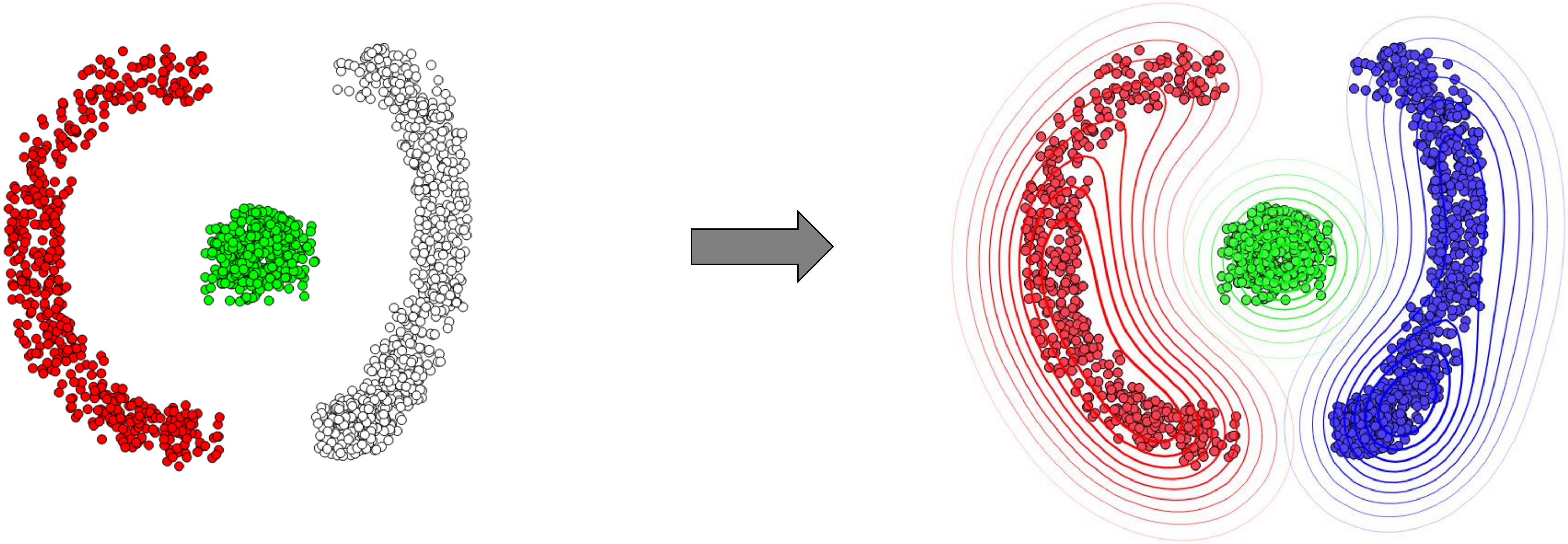
Kernel PCA



Polynomial Kernel

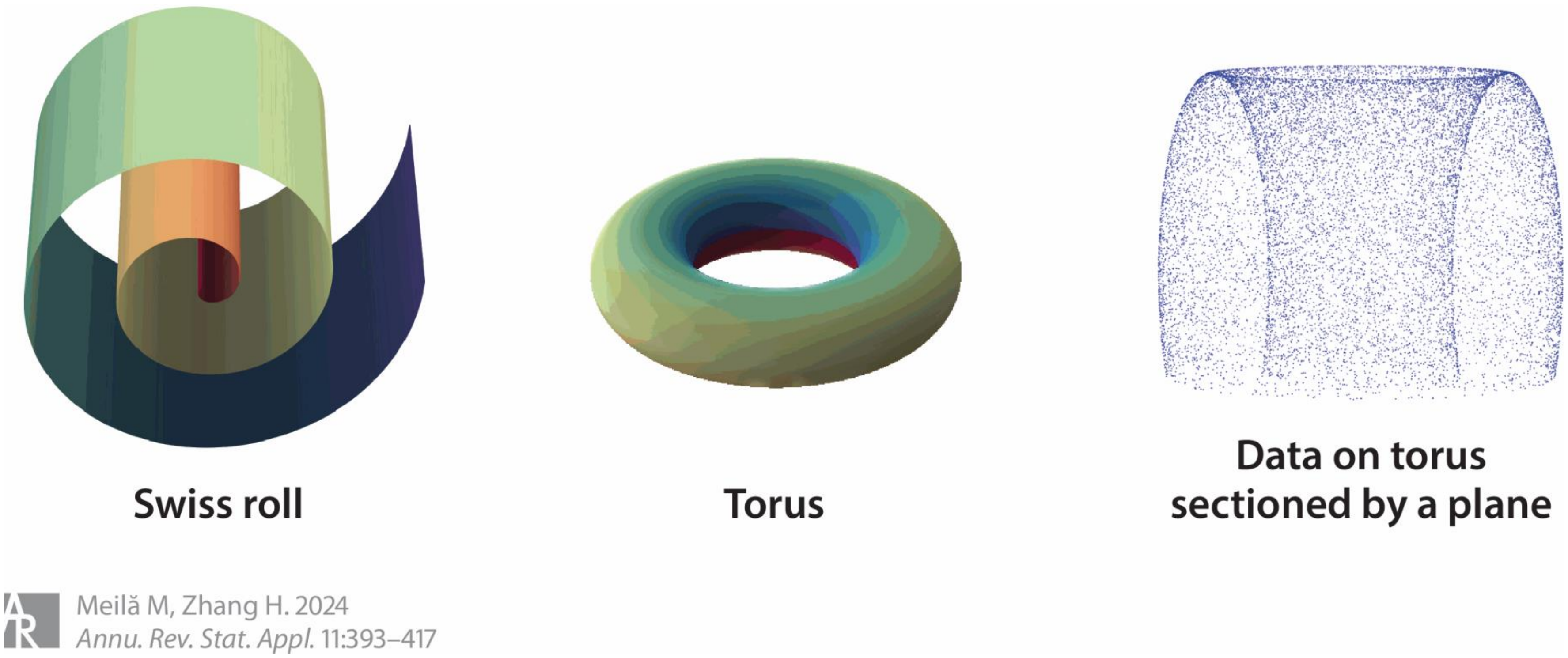
Data becomes linearly separable

Kernel K-means



Can identify non-convex (non-globular) clusters

Manifold Learning



Swiss roll

Torus

Data on torus sectioned by a plane

Meilă M, Zhang H. 2024
Annu. Rev. Stat. Appl. 11:393–417

Figure 2 Examples of manifolds with intrinsic dimension d = 2. (Left) A Swiss roll. (Middle) A torus (hollow). (Right) 1,000 points sampled from a torus sectioned by a plane.

Dataset resides on an unknown nonlinear manifold.

Manifold Learning

Principle:

Dataset resides on an unknown nonlinear manifold.

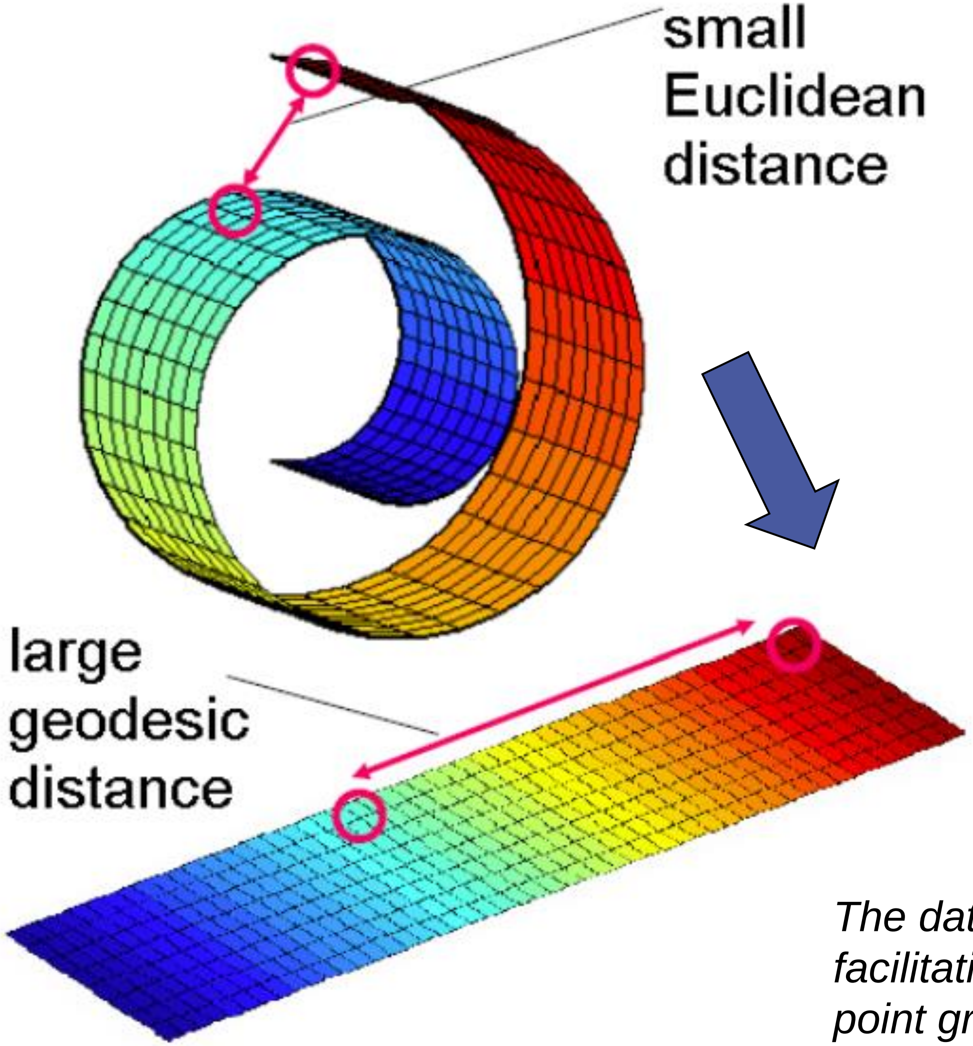
Identify the manifold and construct a new representation.

Project the data onto this representation to simplify further computations.

Manifold Learning

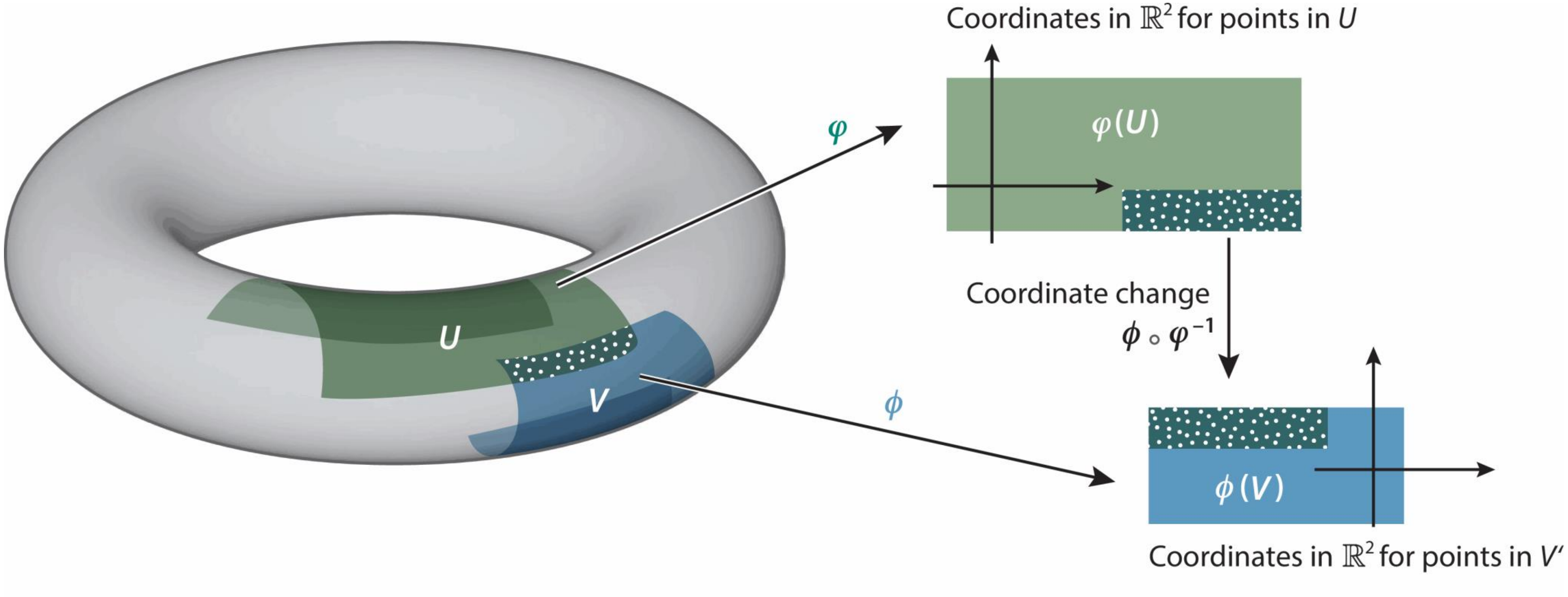
Principle:

- Identify the manifold and construct a new representation.
- Project the data onto this representation to simplify further computations.



The data appears linear after projection, facilitating color-based classification of data point groups.

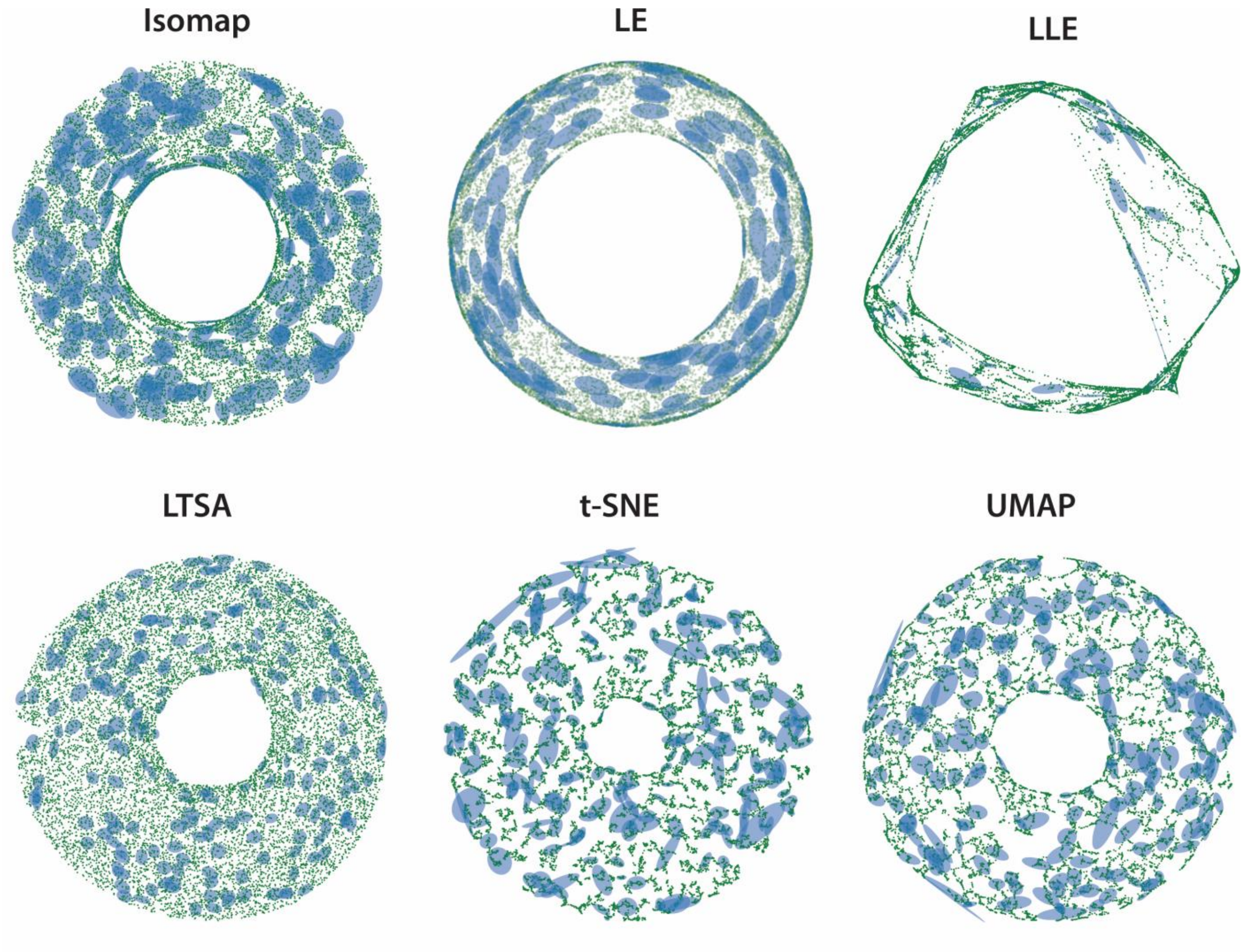
Manifold Learning


Coordinates in $\mathbb{R}^2$ for points in U
φ
$\varphi(U)$
U
V
Coordinate change
$\phi \circ \varphi^{-1}$
ϕ
$\phi(V)$
Coordinates in $\mathbb{R}^2$ for points in V'

Meilă M, Zhang H. 2024
Annu. Rev. Stat. Appl. 11:393–417

Manifold Learning Methods: **Trade-offs**

Sensitivity and properties of learned representations depending on parameters' and algorithmic choices.



Meilă M, Zhang H. 2024
Annu. Rev. Stat. Appl. 11:393–417

Course Overview

- Kernels: definition & role in modifying measures of distance
- Kernel methods:
  - Kernel PCA
  - Kernel CCA
  - Kernel K-means
  - Spectral Clustering
  - Manifold Learning
  - Support Vector Machine
  - Gaussian Process

Unsupervised Learning Methods

Unsupervised, semi-supervised and supervised learning

Course Overview

Kernels: definition & role in modifying measures of distance

Kernel methods:

- Kernel PCA
- Kernel CCA
- Kernel K-means
- Spectral Clustering
- Manifold Learning
- Support Vector Machine
- Gaussian Process

- Based on ***eigendecomposition*** *of a* ***similarity matrix***
- ***Differ*** in the selection of the ***similarity matrix***
- ***Eigenvectors / eigenvalues*** reveal ***data structure***

Prerequisites

We expect you be familiar with the following Machine Learning (ML) Methods:

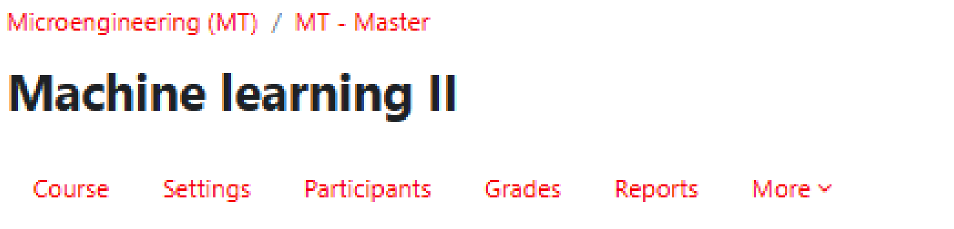
- Principal Component Analysis
- Clustering with K-means and Gaussian Mixture Models
- SVM for classification and regression
- Linear and weighted regression
- Reinforcement learning
- Neural networks

We expect you be familiar with proper ways in which to evaluate ML methods:

- Cross-validation
- F-measure
- ROC curve
- Accuracy, precision, etc.

Class Repositories

Moodle



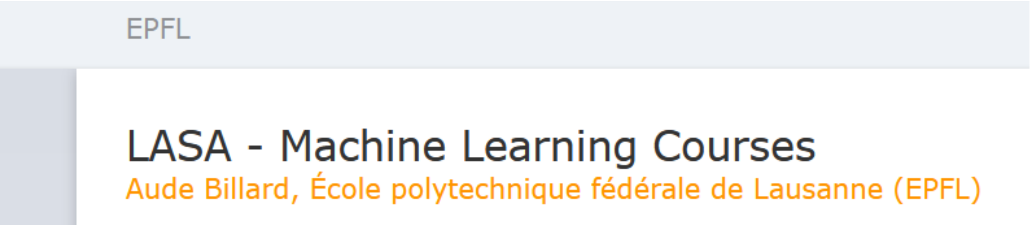
Microengineering (MT) / MT - Master

Machine learning II

https://moodle.epfl.ch/course/view.php?id=14885

- Videos streaming
- Slides
- Exercises & solutions
- Programming exercises & solutions
- Lecture Notes

Video Media Channel



LASA - Machine Learning Courses

Aude Billard, École polytechnique fédérale de Lausanne (EPFL)

https://mediaspace.epfl.ch/channel/LASA+-+Machine+Learning+Courses/30562

Complete repository
of all videos – for Machine Learning I & II

Course Format & Schedule

- Class takes place only on Wednesdays: 13h15-16h00
- **Split format**:
  - ***Watch video of lecture material prior to coming to class! – 45 minutes***
  - 13h15-15h00: Interactive Lecture + interactive exercise session
  - 15h15-16h00: Additional Exercises
- 4 times during semester for all 3 hours (13h15-16h00) - Practice Sessions on computer & Q&A on coding projects - **Check class timetable on moodle!**

Grading Scheme

40% of the grade based on work performed during the semester.

Work is done either alone or in team of two people

Choice between:

1. **Code competition**: analyse a dataset among a list using one on more algorithms seen in class – extract insightful information (inferring patterns, identifying key features, discarding irrelevant data); write a 2-4 pages report and present your analysis in class – **the most insightful analysis wins!**

OR

2. **Debates:** choose a topic and your camp (pro or con) among a list of topics - read papers on the chosen topic; – write a 2-4 pages argumenta and present your arguments in a live panel/debate held in class - **the most convincing debating team wins!**

60% based on final oral exam

Closed book, but you are allowed to bring a recto-verso A4 page with handwritten personal notes

Topics for Debates:

Success of ML depends most on :

a) the data;

b) the algorithm;

c) none of these

Which approach is the most effective for real-world applications?

a) Supervised learning;

b) Unsupervised Learning

Real-world deployment will remain hindered until we have:

a) better methods for incremental learning;

b) good interfaces for lay users;

c) more efficient storage methods and faster computing resources

Generalisation in machine learning:

a) can be readily assessed with several metrics and benchmarks;

b) is ubiquitous;

c) is an ill-posed problem

Biases in machine learning :

a) can be resolved with heavier computation & curated datasets;

b) is not an issue for the vast majority of applications;

c) is by essence irresolvable

Machine learning and climate change:

a) growth of the former is incompatible with the latter,

b) both can be made compatible under certain constraints;

c) this is a non-issue

Debates Guidelines: Machine Learning Tradeoffs

The goal of these debates is to engage in an in-depth discussion on controversial topics related to machine learning techniques and their real-world applications.

Preparation Requirements:

- **Research:** Read relevant literature to build a well-informed perspective.
- **Materials:** Prepare a slide deck and a 2-4 pages summary outlining your key arguments.

Debate Structure:

1. Opening Statements:

- Each debater presents their position using 1-2 slides.

2. Live Debate (15-20 minutes):

- Participants engage in discussion, debating with one another and responding to audience questions.
- Debaters must support their arguments with additional slides, providing concrete examples, quantifiable results, etc.

The debates are an opportunity to critically analyze different viewpoints, defend your position with well-researched arguments, and learn on how to engage in a dynamic discussion and build convincing arguments.

Coding Competitions - Topics

Average Monthly surface temperature (1940-2024)

https://www.kaggle.com/datasets/samithsachidanandan/average-monthly-surface-temperature-1940-2024

This Dataset contains details of Average Monthly surface temperature (1940-2024). Current climate change is primarily caused by human emissions of greenhouse gases. This warming can drive large changes in sea level, sea ice and glacier balances, rainfall patterns, and extreme temperatures. This has potentially devastating impacts on human health, farming systems, the stability of societies, and other species.

Coding Competitions - Topics

Sleep Health and Digital Screen Exposure Dataset

https://www.kaggle.com/datasets/arifmia/sleep-health-and-digital-screen-exposure-dataset

This dataset contains multiple health-related attributes collected from individuals, including their sleep quality, stress levels, heart rate, and screen exposure habits. It can be used for statistical analysis, machine learning modeling, and health-related research.

Coding Competitions - Topics

- **Full Netflix Dataset**

https://www.kaggle.com/datasets/octopusteam/full-netflix-dataset/data

This dataset provides a comprehensive collection of all titles (Movies and TV Series) available on Netflix. In addition to basic information, it includes IMDb-specific data like IMDb ID, Average Rating, and Number of Votes.

Coding Competitions - Topics

- **Traffic Accidents**

  https://www.kaggle.com/datasets/oktayrdeki/traffic-accidents

This dataset contains detailed information on traffic accidents across various regions and time periods. It includes various metrics such as accident date, weather conditions, lighting conditions, crash types, injuries, and vehicle involvement. The data span multiple locations and accident types, offering a comprehensive view of traffic incidents and their causes.

Coding Competitions - Topics

Food Nutrition Dataset

https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset/data

The dataset provides detailed nutritional information for a wide range of food items commonly consumed around the world. This dataset aims to support dietary planning, nutritional analysis, and educational purposes by providing extensive data on the macro and micronutrient content of foods.

Coding Competitions - Topics

Calories Burned During Exercise and Activities

https://www.kaggle.com/datasets/aadhavvignesh/calories-burned-during-exercise-and-activities/data

This dataset contains the number of calories burned by a person while performing some activity/exercise. It contains 248 activities and exercises ranging from running, cycling, calisthenics, etc.

Coding Competitions - Topics

AI/ML Salaries

https://www.kaggle.com/datasets/cedricaubin/ai-ml-salaries/data

The salaries are from ai-jobs. Ai-jobs collects salary information anonymously from professionals all over the world in the AI/ML and Big Data space and makes it publicly available for anyone to use, share and play around with. The data is being updated regularly with new data coming in, usually on a weekly basis.

Coding Competitions - Topics

Education & Career Success

https://www.kaggle.com/datasets/adilshamim8/education-and-career-success

This dataset explores the relationship between academic performance and career success. It includes 5000 records of students' educational backgrounds, skills, and career outcomes. The dataset can be used for various analyses, such as predicting job success based on education, identifying key factors influencing salaries, and understanding the role of networking and internships in career growth.

Coding Competition Guidelines: Machine Learning for Structure Discovery

The objective of this competition is to apply machine learning techniques for structure discovery, as covered in class, to a real-world dataset.

Expectations:

1. Initial Data Analysis:

- Perform a preliminary analysis of the dataset using standard statistical methods.
- Visualize data distributions and compute key statistics (e.g., mean, median).

2. Application of Machine Learning Techniques:

- Implement at least one (or multiple) structure discovery techniques from class.
- You may also explore alternative approaches, such as manifold learning, kernel-based clustering, or dimensionality reduction methods.

3. Identifying Key Insights:

- Evaluate the results and determine the most meaningful findings.
- Prepare a 2-4 pages report and accompanying slides that address the following:
    - Techniques Used: Justify your choice of method(s).
    - Hyperparameter Selection: Explain how you tuned the model’s parameters.
    - Findings & Insights: Analyze the results and their implications.
    - Limitations & Challenges: Discuss any shortcomings or potential biases in your analysis.

The goal is not only to apply machine learning techniques but also to critically assess their effectiveness and limitations in uncovering meaningful structure within the dataset.