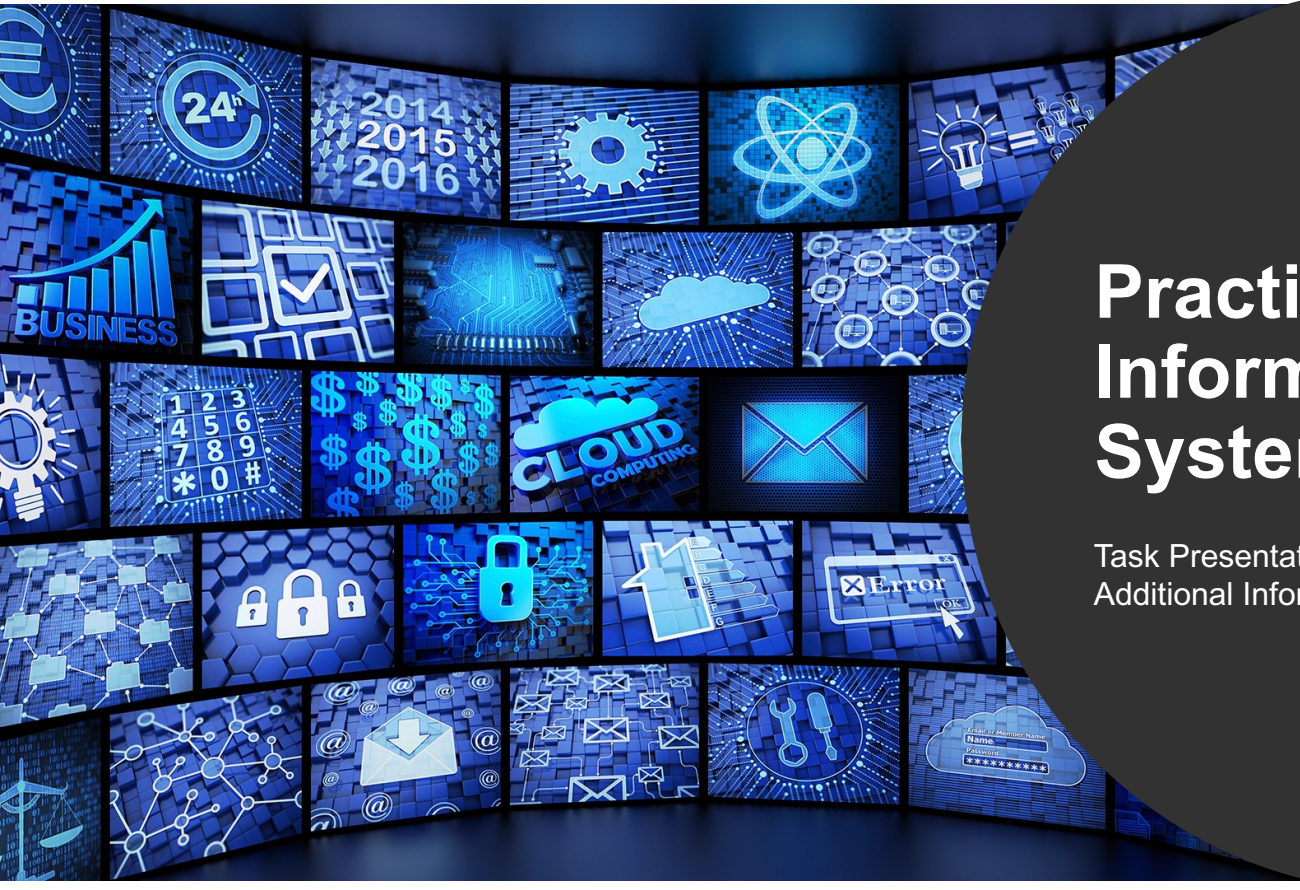
University of Stuttgart
Germany



Practical Course Information Systems

Task Presentations, Team Formation & Additional Information

Winter Term 2023/24

Introduction

- Topic: Data-intensive computing and analytics

- Task: Use or implement a new or extend an existing data-intensive application based on tasks that may involve data analysis, data visualization, and efficient data processing

- Goals:
  - application development in small teams
  - learn about systems for large-scale data processing
  - recent, relevant research topics

General Information

- Official start: right **NOW** ☺
- End: April 2024 (final presentation)

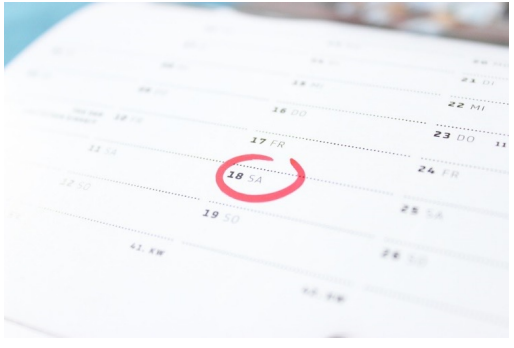
- Scope: 6 CP (graded) → **180 hours/person**

- Grading Criteria:
  - Implementation (completeness, quality, …)
  - General impression (reliability, development approach, …)
  - Final presentation

General Information

- You may deregister from this course during the next **two weeks**
- To do so, write an e-mail to your supervisor, to Yunxuan Li (yunxuan.li@ipvs.uni-stuttgart.de), **and** to Professor Holger Schwarz (holger.schwarz@ipvs.uni-stuttgart.de)

- Online **exam registration** (C@mpus) necessary
- Note the **deadline** of the examination office



What we expect …

- Basic knowledge of database and information systems, e.g., from the course “Modellierung”
- Programming skills in at least one programming language
- Readiness for teamwork
- Ability to work independently and self-organized

Topic Presentations

- **3 persons** per developer team → 4 teams

- Each team works on one of the following topics:
  - Topic 1: Extension of AdaPrivFlow
  - Topic 2: Privacy Protection of Trajectory Data
  - Topic 3: Automated Machine Learning for Clustering Analyses
  - Topic 4: Automated Machine Learning for Classification Ensembles

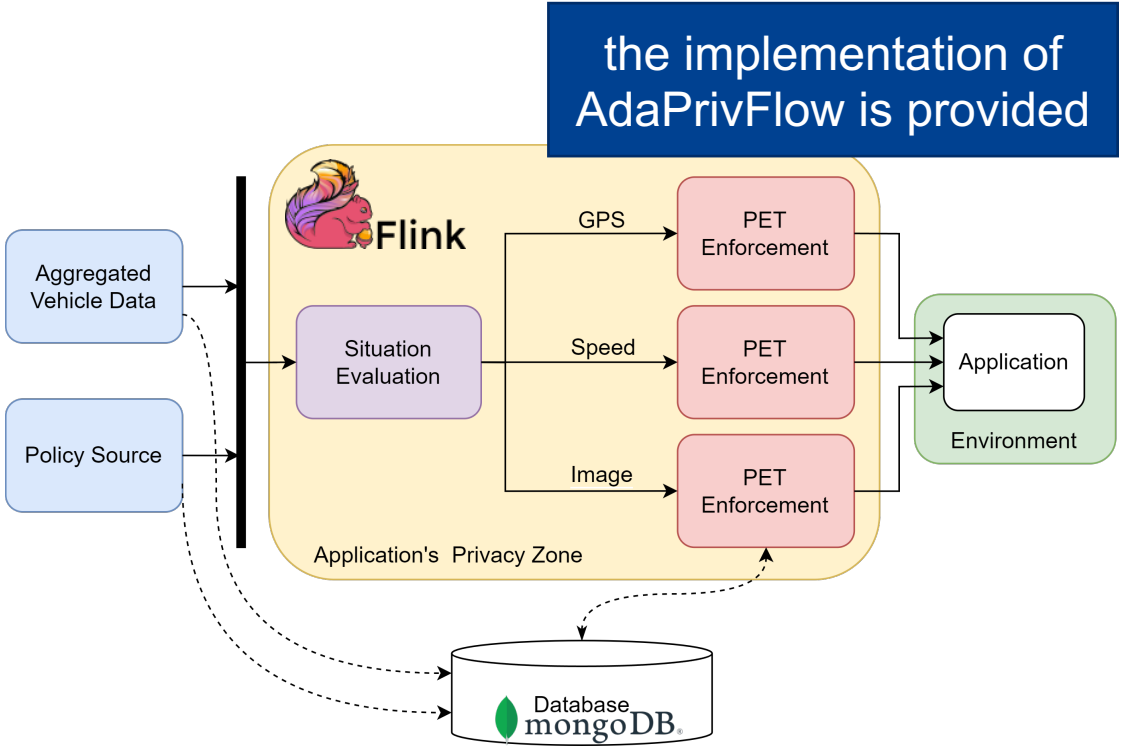
Topic 1: Extension of AdaPrivFlow

Yunxuan Li

Extension of AdaPrivFlow

Motivation: Situation-Aware Privacy-Protection for Connected Vehicles

- Connected Vehicles can gather, process, and share data, which raises **privacy concerns**
- Privacy protection for Connected Vehicles must consider the **dynamic and context-dependent** nature of drivers' privacy needs



Extension of AdaPrivFlow

Main Tasks

- Extend / improve the AdaPrivFlow with following points:
  - SubTask 1: use CARLA as the data generator
  - SubTask 2: extend the pipeline to support Lidar data
  - SubTask 3: break up the pipeline component with Kafka
- Explore the possibility of deploying the AdaPrivFlow pipeline into the Docker

Extension of AdaPrivFlow

SubTask 1: use CARLA as data generation

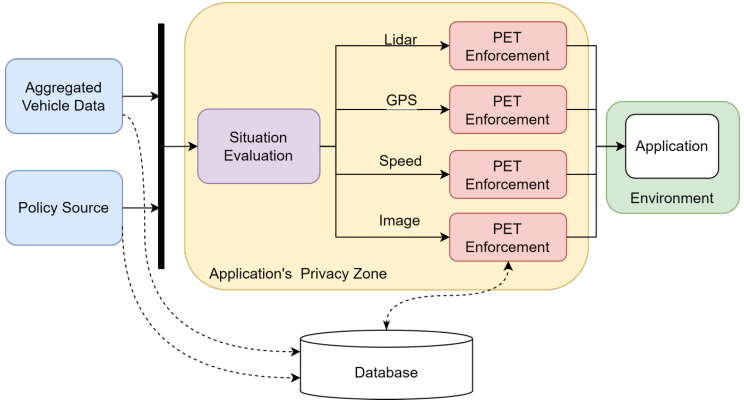
- CARLA Simulator
  - an open-source simulator for autonomous driving research
  - CARLA boasts an impressive array of models of real-world sensors like cameras, LIDAR and RADAR
  - allows users to define and execute different traffic situations based on modular behaviors
- Goal:
  - retrieve data from a simulated car within Carla
  - prepare the data for AdaPrivFlow
  - send data to AdaPrivFlow

Extension of AdaPrivFlow

SubTask 2: extend the pipeline to support Lidar data

- **Li**ght **D**etection **a**nd **R**anging Sensor (Lidar)
  - uses a laser to measure distances
  - to capture very detailed and wide-ranging information about a vehicle's surroundings
  - Lidar data is collected as a "point cloud"

- Goal:
  - explore different data types to store and transfer Lidar data
  - extend AdaPrivFlow to support Lidar data

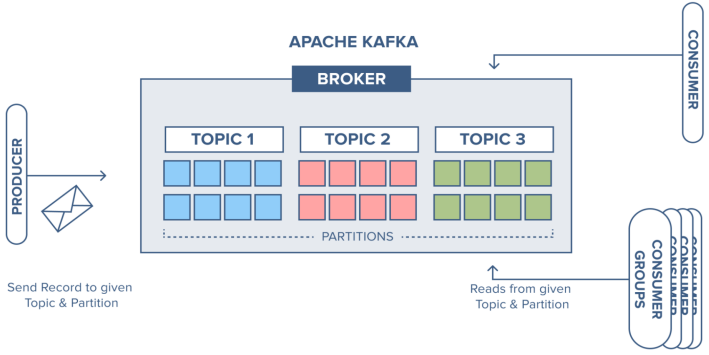


picture source: https://carla.readthedocs.io/en/latest/ref_sensors/#lidar-sensor

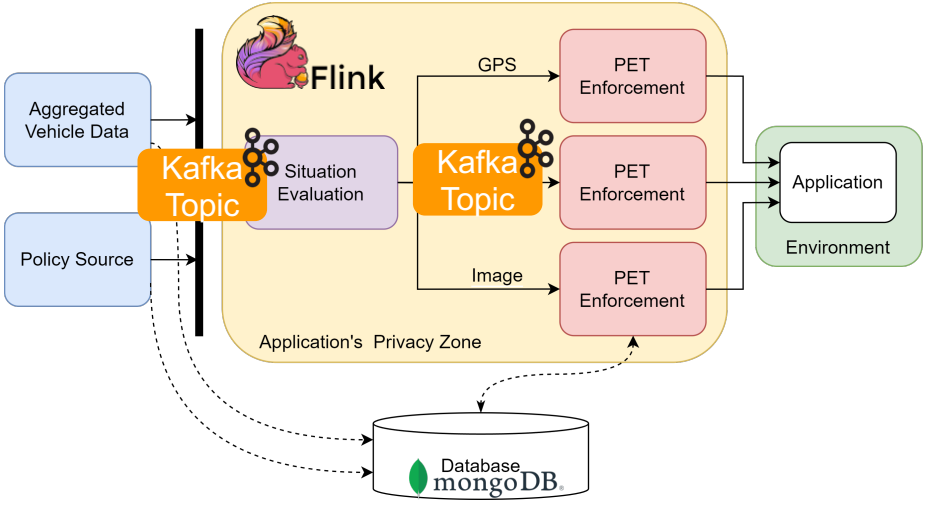
Extension of AdaPrivFlow

SubTask 3: break up the pipeline component with Kafka

- Kafka
  - initially conceived as a messaging queue
  - allow to publish and subscribe to streams of events through **topics**



- Goal:
  - adding corresponding Kafka topics to the data pipeline



Task Summary

- Research and familiarize with the required techniques
- Complete the three subtasks
- Explore the possibility of deploying the AdaPrivFlow pipeline into the Docker
- Write the documentation
- Prepare a final presentation

Topic 2: Privacy Protection of Trajectory Data

Yunxuan Li

Extension of AdaPrivFlow

Motivation: Privacy-Protection for Trajectory Data

- Continuously sharing trajectory data may end up revealing a great amount of information in terms of a user's **behavior, mobility patterns, and social relationships**
- Four types of methods to protect trajectory data:
  - spatial-location cloaking approaches
  - temporal cloaking methods
  - addition of redundant dummy locations
  - path confusion

Extension of AdaPrivFlow

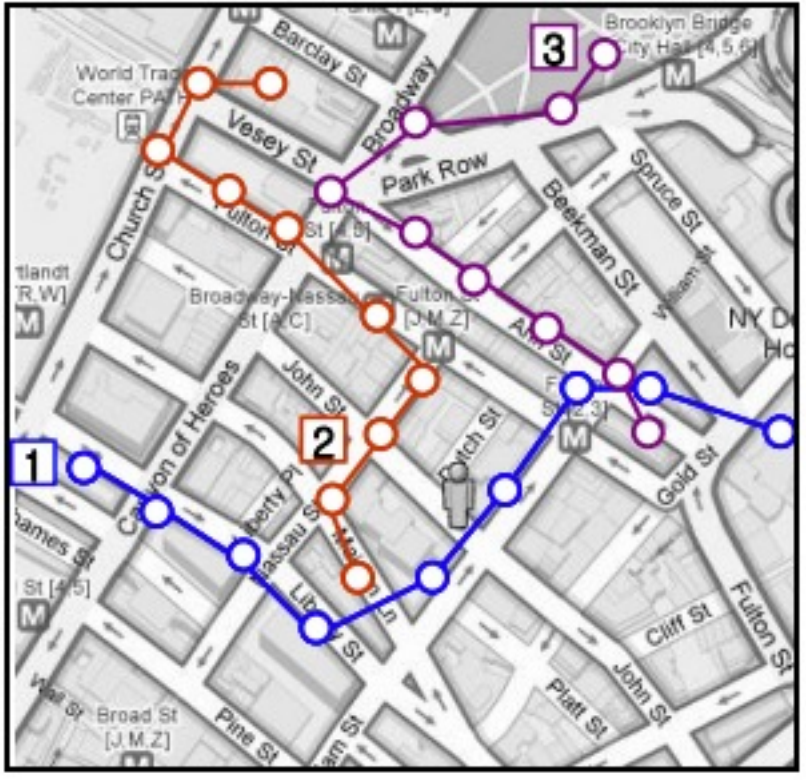
Main Tasks

- spatial-location cloaking approaches
  - the exact location of the user is replaced by a broader spatial region termed cloaking region
- temporal cloaking methods
  - uses time transformation and delays the user's response by a time period
- addition of redundant dummy locations
  - instead of sharing the actual location, share one or more locations that are very close to the actual one (generated dummy locations)
- path confusion
  - use a perturbation technique to change the actual trajectory of a user (i.e., publish another report instead of the actual one)

Extension of AdaPrivFlow

Main Tasks

- Data Source
  - use the trajectory data simulated by Carla
- Result Visualisation
  - explore the possibility of demonstrating the generated trajectory data with Carla
  - if it is not possible: implement a Dashboard to visualise the result



Task Summary

- Research and familiarize with the four given methods
- Implement these methods
- Test and compare these methods with trajectory data simulated by Carla
- Visualise the resulting trajectory of each implemented method
- Write the documentation
- Prepare a final presentation

Task 3/4: Automated Machine Learning for Clustering Analyses / Classification Ensembles

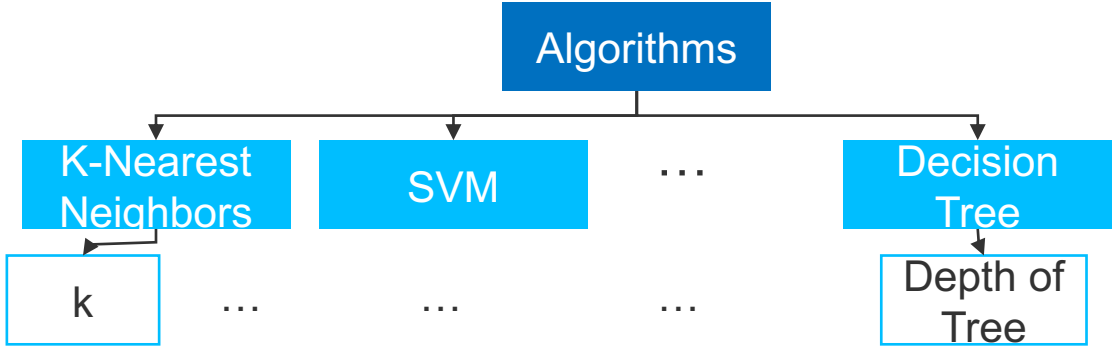
Dennis Treder-Tschechlov, Julius Voggesberger

Motivation

- Machine Learning (ML) used in many application domains
  - IoT, Advertisement/Marketing, Medicine, Language support, …
- ML Algorithm and Parameters are dependent on concrete use case

→ Which algorithms?

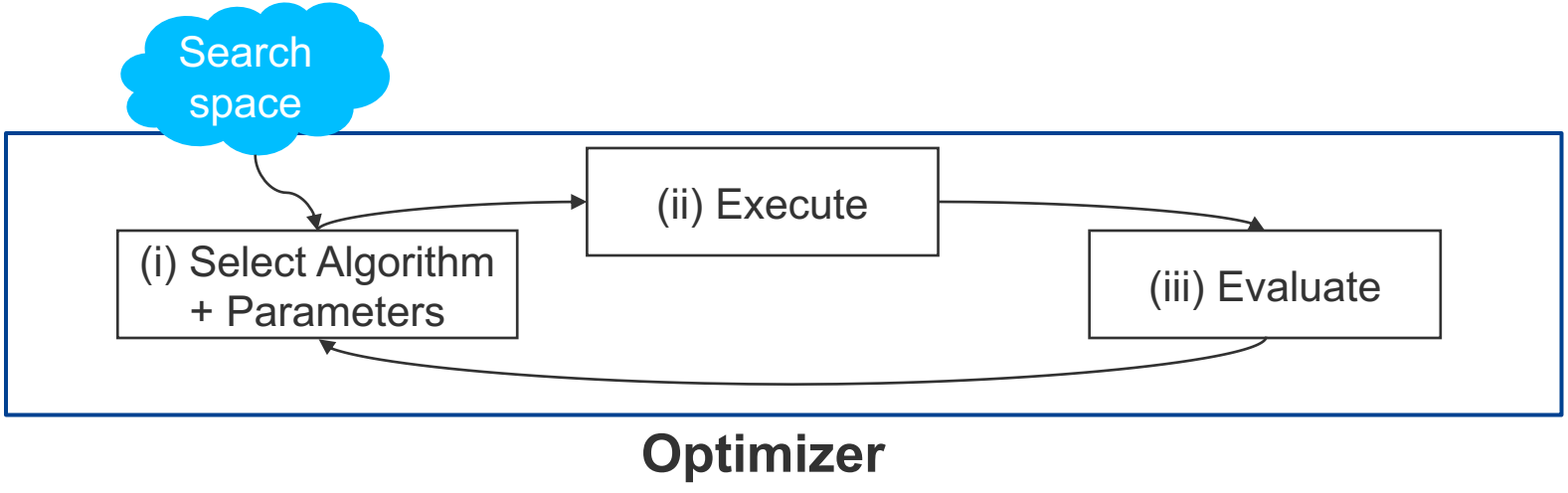
→ Which parameters?



Solution: Automated Machine Learning (AutoML)

Automated Machine Learning (AutoML)

- Given a budget (e.g., one hour) find a solution that is *as accurate as possible*
- Very Large search spaces → Optimizer to find good solutions



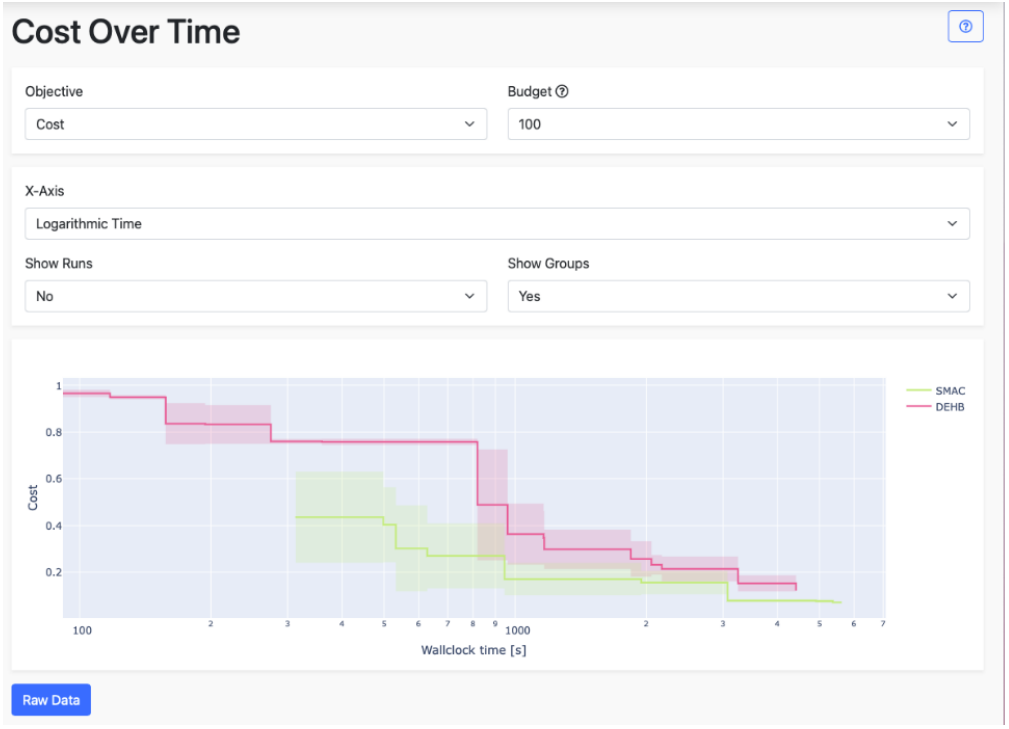
  - Optimizer: Random Search, Grid Search, Bayesian Optimization, …

Optimization Library

- Sequential Model-Based Algorithm Configuration (**SMAC**)
  - https://github.com/automl/SMAC3
- Supports different optimizers → easily interchangeable
- User has to …
  - define search space
  - write target evaluation function
  - parse results
- Definition of search space w. *ConfigSpace*
  - https://github.com/automl/ConfigSpace

Tasks

- Update to SMAC V2
  - Extend search space
  - Add new optimizers
  - Adapt interface for optimizers
  - Parsing of (intermediate-)results
- UI (e.g. using DeepCave) for …
  - tracking optimization results
  - Defining search space
  - Defining input parameters



Task 3 / 4 Organization

We provide implementations of (old) SMAC versions for the tasks

- **Task 3: AutoML for Clustering Analyses (AutoML4Clust)**
  - Goal: Find *natural* groupings(=clusters) in the data
- **Task 4: AutoML for Classification Ensembles (AutoML-OpEn)**
  - Goal: Combine results of different classification models

- In the first 1-2 months:
  - Each group familiarizes themselves with their task
  - exchange knowledge/ discuss problems

Organization

Topic Presentations

- **3 persons** per developer team → 4 teams

- Each team works on one of the following topics:
  - Topic 1: Extension of AdaPrivFlow
  - Topic 2: Privacy Protection of Trajectory Data
  - Topic 3: Automated Machine Learning for Clustering Analyses
  - Topic 4: Automated Machine Learning for Classification Ensembles

Participants and Team Formation

Topic	Team Members	Advisor
1	Agnihotri Anusha, Chandra Rahul, Yadav Sachin	Yunxuan L.
2	Bartels Malte Henrik, Pisano Vincenzo, Fan Tingyu	Yunxuan L.
3	Abouhelal Hassan, Mohamed Mohamed, Baba Malek	Dennis TT.
4	Hann Kilian, Xie Siwei, Saaran Viraat(?)	Julius V.

Next Steps

1. Arrange a meeting with your advisors via e-mail within this week
   - Yunxuan: yunxuan.li@ipvs.uni-stuttgart.de
   - Dennis: dennis.tschechlov@ipvs.uni-stuttgart.de
   - Julius: julius.voggesberger@ipvs.uni-stuttgart.de

2. Familiarization with the topic
   - Get topic-specific literature and additional information

Questions?