CIVIL-557

# Decision-aid methodologies in transportation

## Lecture 5: Issues with performance validation

Tim Hillel

Transport and Mobility Laboratory TRANSP-OR
École Polytechnique Fédérale de Lausanne EPFL

# Last week

- Ensemble method theory
  - Bagging (bootstrap aggregating) and boosting
  - Random Forest
  - Gradient Boosting (XGBoost)

- Hyperparameter selection theory
  - $k$-fold Cross-Validation
  - Grid search

# Today

1. Homework feedback/recap
2. Hierarchical data and grouped sampling
3. Advanced hyperparameter selection methods
4. Project introduction

# Hyperparameter selection homework

Discussion of worked example

# Performance estimate discrepancy

## Cross-validation

- Train on 4 folds, test on 1 fold
  - Training data: 80% of train-validate data
- Random sampling
  - Internal validation

## Test

- Train on first two years, test on final year
  - Training data: 100% of train-validate data
- Sample by year
  - External validation

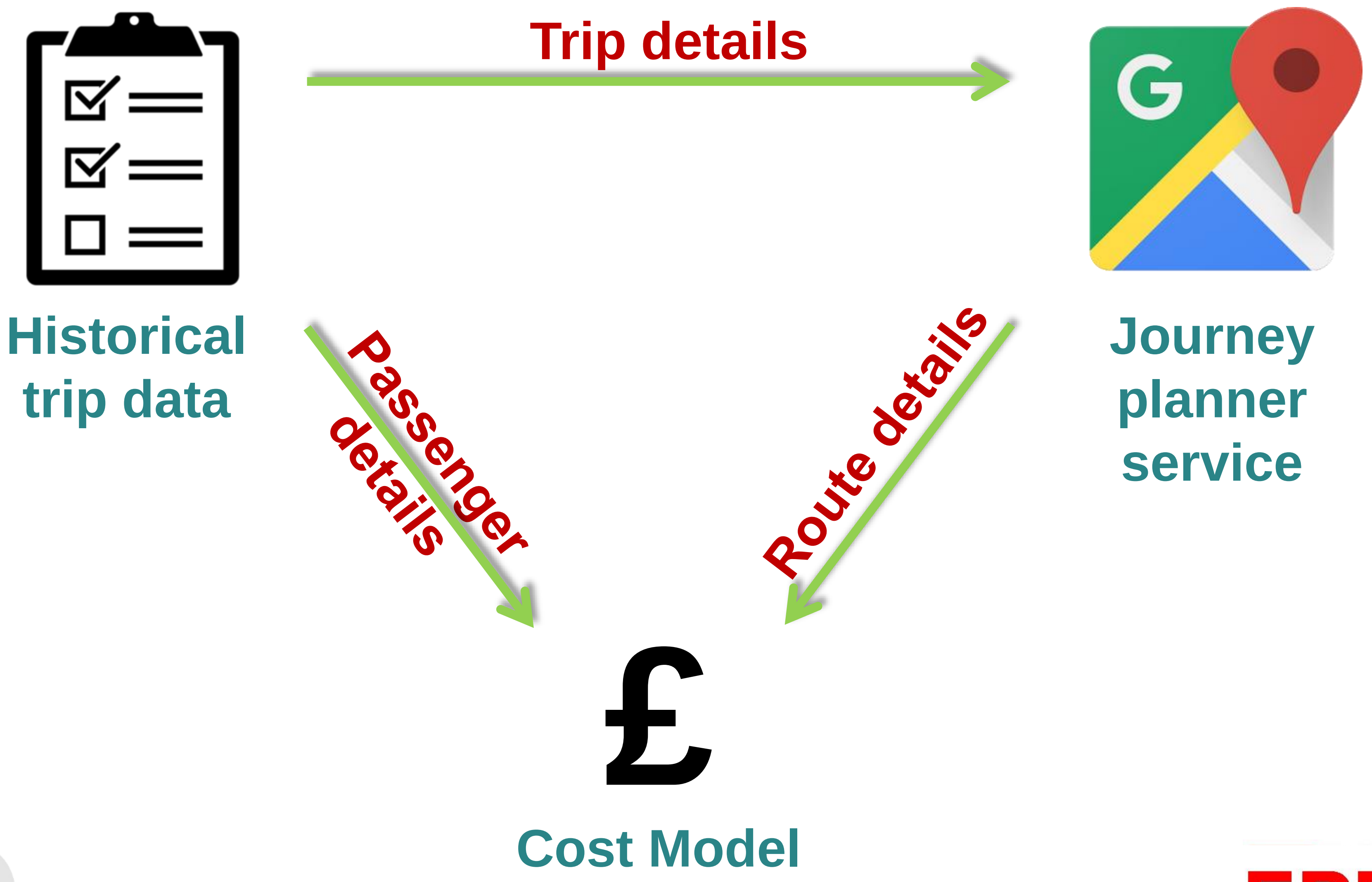
TRANSP-OR

EPFL

# Impacts of random sampling

Why the discrepancy?

# Dataset building process



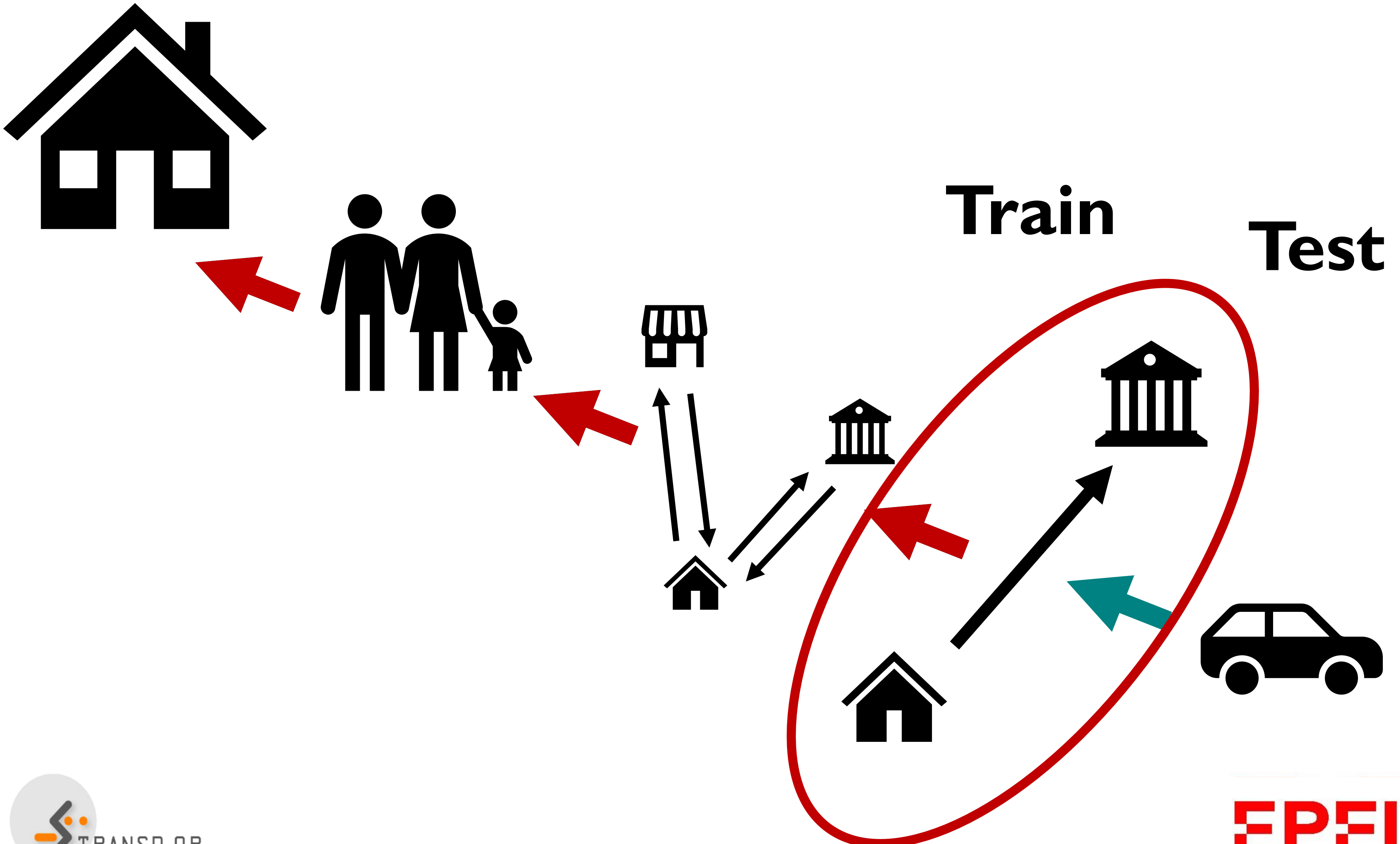
# Dataset building process

**Historical trip data**

London Travel Demand Survey (LTDS)
- Annual rolling **household travel survey**
- Each household member fills in **trip diary**

3 years of data (2012/13-2014/15)
- ~130,000 trips

# Random Sampling



# State of practice

**Systematic review:**

***ML methodologies for mode-choice modelling***

**60 papers → 63 studies**

# State of practice

**56% (35 studies) use hierarchical data**

**All use trip-wise sampling**

# Implications

- Mode choice heavily correlated for return, repeated, and shared trips. E.g.:
  - Return journey to/from work
  - Repeated journey to doctor’s appointment
  - Shared family trip to concert

- Journey can be any combination of return/repeated/shared

# Implications

- Random sampling – return/repeated/shared trips occur across folds
- These trips have some correlated/identical features
  - E.g. trip distance, walking duration, etc
- ML model can recognise unique features and recall mode choice for trip in training data – **data leakage**

# Implications

- Model performance estimate will be optimistically biased using random sampling for hierarchical data

What about selected hyperparameters?

# London dataset

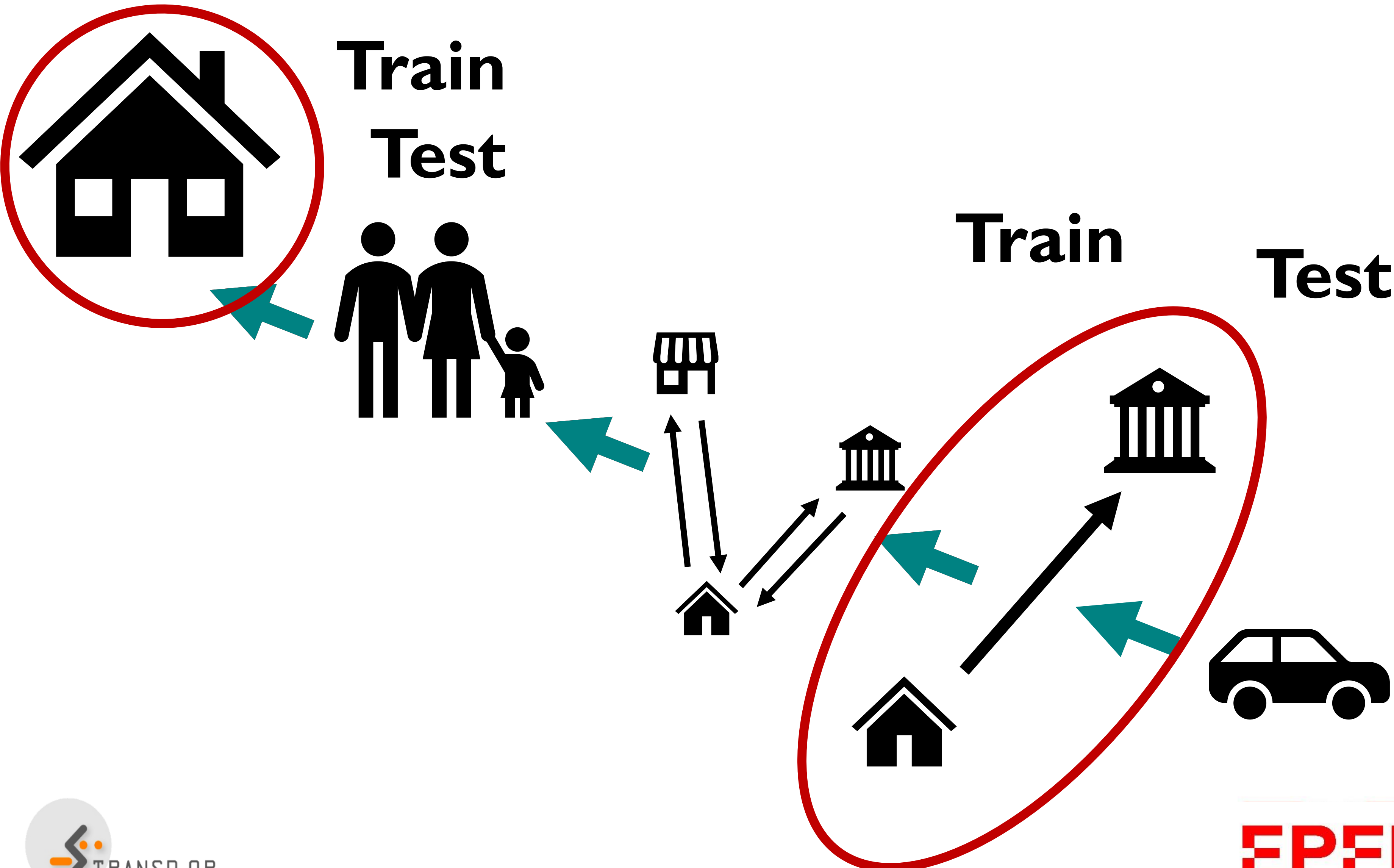
| Type | Pairs/sets | No. Trips | No. Trips matching mode | Proportion matching mode |
|---|---|---|---|---|
| Return | 15 605 | 32 471 | 30 898 | 95.2 % |
| Repeated | 1315 | 2711 | 2496 | 92.1 % |
| Shared | 8541 | 20 623 | 20 051 | 97.2 % |
| **All** | 15 814 | 40 520 | 39 357 | 97.1 % |

74% of trips in training data (first two years) belong to pairs or sets of return/repeated/shared trips

# Trip-wise sampling

| | CV | Test | Diff |
|---|---|---|---|
| **LR** | 0.676 | 0.693 | 0.017 |
| **FFNN** | 0.680 | 0.696 | 0.017 |
| **RF** | 0.545 | 0.679 | 0.134 |
| **ET** | 0.536 | 0.685 | 0.149 |
| **GBDT** | 0.467 | 0.730 | 0.263 |
| **SVM** | 0.579 | 0.823 | 0.244 |

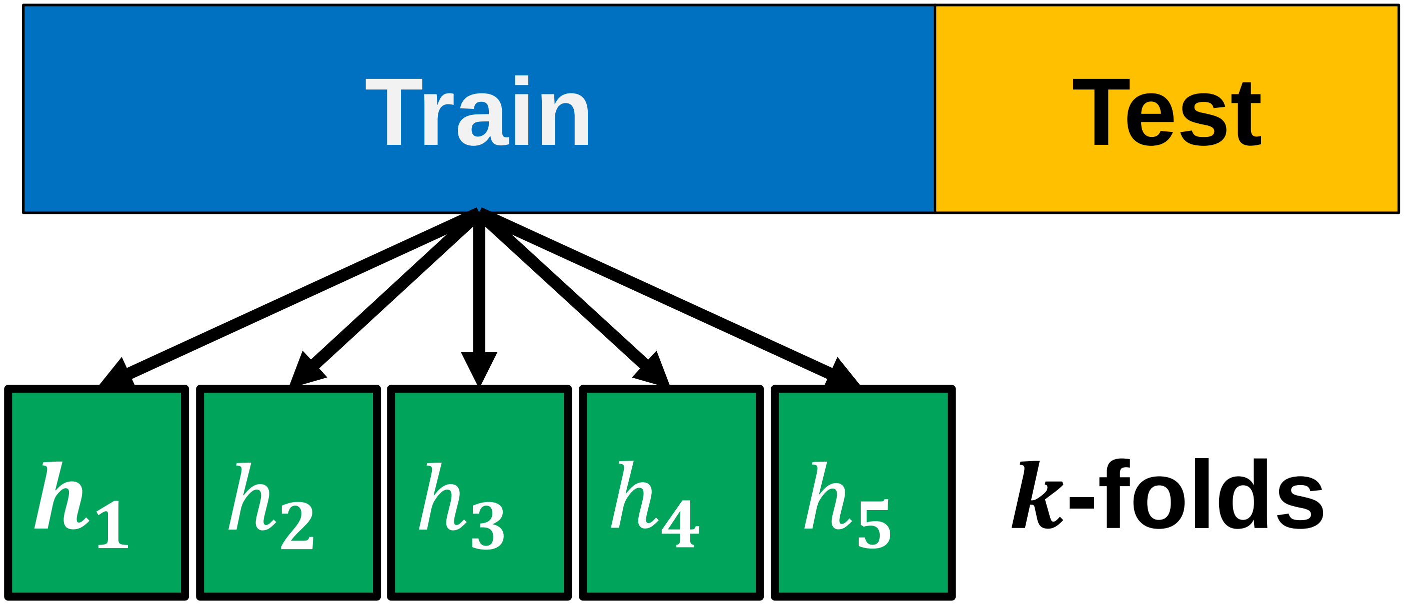
# Solution - Grouped Sampling



# Solution – grouped sampling

- Trips by one household appear purely in single fold
- Prevents data leakage from return/repeated/shared trips

# Grouped cross-validation




Sample by household index into groups $h_i$

# Trip-wise sampling

| | CV | Test | Diff |
|---|---|---|---|
| **LR** | 0.676 | 0.693 | 0.017 |
| **FFNN** | 0.680 | 0.696 | 0.017 |
| **RF** | 0.545 | 0.679 | 0.134 |
| **ET** | 0.536 | 0.685 | 0.149 |
| **GBDT** | 0.467 | 0.730 | 0.263 |
| **SVM** | 0.579 | 0.823 | 0.244 |

# Grouped sampling

| | CV | Test | Diff |
|---|---|---|---|
| **LR** | 0.679 | 0.693 | 0.014 |
| **FFNN** | 0.679 | 0.688 | 0.009 |
| **RF** | 0.656 | 0.677 | 0.021 |
| **ET** | 0.658 | 0.680 | 0.022 |
| **GBDT** | 0.634 | 0.651 | 0.017 |
| **SVM** | 0.679 | 0.692 | 0.013 |

# Hyperparameter selection

Can we beat grid search?

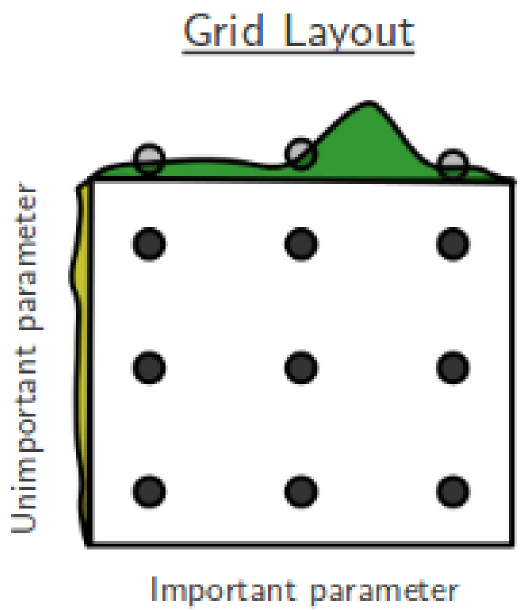
# Grid-search

- Predefine search values for each hyperparameter
- Search all combinations in exhaustive grid-search

- Simple to understand, implement, and parallelise

- Inefficient:
  - Lots of time evaluating options which are likely to be low performing
  - Few unique values for each hyperparameter tested

# Grid search




***Random Search for Hyper-Parameter Optimization*, Bergstra et al (2012)**

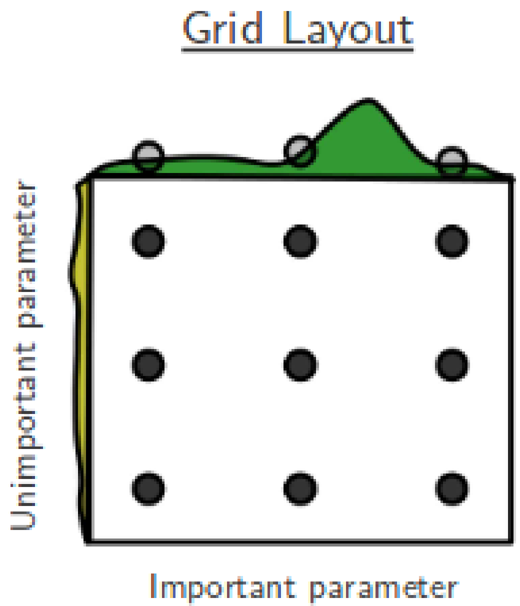
# Advanced hyperparameter selection

- Other alternatives to grid-search:
  - Random search
  - Sequential Model Based Estimation (SMBO)

# Random search

- Define search distributions for each hyperparameter
  - E.g. uniform integer between 1-50 for max-depth
  - Can be binary, normal, lognormal, uniform, etc
- Simply draw randomly from distributions from each distribution

# Random search




***Random Search for Hyper-Parameter Optimization*, Bergstra et al (2012)**

# Random search

- Unique values for each iteration for each hyperparameter
- Even easier to parallelise than grid-search!
- Outperforms grid-search in practice
- However, still wastes time evaluating options which are likely to be low performing

# SMBO

- As with random search, define search distributions for each hyperparameter

- However, base sequential draws on previous results
  - Lower likelihood of choosing values close to others which perform poorly
  - Higher likelihood of choosing values close to others which perform well

# SMBO

- Several algorithms for sequential search
  - Gaussian Processes (GP)
  - Tree-structured Parzen Estimator (TPE)
  - Sequential Model-based Algorithm Configuration (SMAC)
  - …

- Several available libraries in Python
  - hyperopt, spearmint, PyBO

# Q&A

- Questions from any part of the course material?

Further Q&A on May 28th

TRANSP-OR

EPFL

# Hands on

## Notebook 1: Advanced hyperparameter selection