

# Automatic machine learning

Tadej Štajner<sup>1</sup>

29.05.2015 / PyData Berlin Lightning talk

**here**

# What is it

- A lot of time in data science engagement and research is spent on model selection
- This involve tons of work tuning knobs and black magic.
- `[SVC(C=c) for c in np.logspace(-4,4,10)]`

# What is it

- A lot of time in data science engagement and research is spent on model selection
- This involve tons of work tuning knobs and black magic.
- `[SVC(C=c) for c in np.logspace(-4,4,10)]`

# What is it

- A lot of time in data science engagement and research is spent on model selection
- This involve tons of work tuning knobs and black magic.
- `[SVC(C=c) for c in np.logspace(-4, 4, 10)]`

# Can we automate this?

There's a pretty systematic pattern:

- Figure out initial baseline models and parameters
- Tune various model parameters until things work better

# Can we automate this?

There's a pretty systematic pattern:

- Figure out initial baseline models and parameters
- Tune various model parameters until things work better

# Can we automate this?

There's a pretty systematic pattern:

- Figure out initial baseline models and parameters
- Tune various model parameters until things work better

## Current state

- **scikit-learn has GridSearch and RandomizedSearch, plus \*CV versions with built-in optimization paths**
- Active research in hyper-parameter optimization - `hyperopt`, `hyperopt-sklearn`



## Current state

- scikit-learn has `GridSearch` and `RandomizedSearch`, plus \*CV versions with built-in optimization paths
- Active research in hyper-parameter optimization - `hyperopt`, `hyperopt-sklearn`

# AutoML Competition

- A multi-stage competition for automatic machine learning approaches on diverse datasets
- `automl.org`
- `codalab.org/competitions/2321`
- Several iterations of tuning phase + new data release phase
- Upload your model code, see where you rank

# AutoML Competition

- A multi-stage competition for automatic machine learning approaches on diverse datasets
- `automl.org`
- `codalab.org/competitions/2321`
- Several iterations of tuning phase + new data release phase
- Upload your model code, see where you rank

# AutoML Competition

- A multi-stage competition for automatic machine learning approaches on diverse datasets
- `automl.org`
- `codalab.org/competitions/2321`
- Several iterations of tuning phase + new data release phase
- Upload your model code, see where you rank

# AutoML Competition

- A multi-stage competition for automatic machine learning approaches on diverse datasets
- `automl.org`
- `codalab.org/competitions/2321`
- Several iterations of tuning phase + new data release phase
- Upload your model code, see where you rank

# AutoML Competition

- A multi-stage competition for automatic machine learning approaches on diverse datasets
- `automl.org`
- `codalab.org/competitions/2321`
- Several iterations of tuning phase + new data release phase
- Upload your model code, see where you rank

# What I am doing

- Design sensible parameter spaces and preprocessing layer combinations for various scikit-learn approaches for different ML problems
- My submission - <http://github.com/tadejs/autokit>, based on `hyperopt-sklearn`
- 3rd place on 1st new data release phase of AutoML competition :)

# What I am doing

- Design sensible parameter spaces and preprocessing layer combinations for various scikit-learn approaches for different ML problems
- My submission - <http://github.com/tadejs/autokit>, based on hyperopt-sklearn
- 3rd place on 1st new data release phase of AutoML competition :)



# What I am doing

- Design sensible parameter spaces and preprocessing layer combinations for various scikit-learn approaches for different ML problems
- My submission - <http://github.com/tadejs/autokit>, based on `hyperopt-sklearn`
- 3rd place on 1st new data release phase of AutoML competition :)

| RESULTS |              |          |            |            |            |            |            |              |
|---------|--------------|----------|------------|------------|------------|------------|------------|--------------|
|         | User         | <Rank>   | Set 1      | Set 2      | Set 3      | Set 4      | Set 5      | Duration     |
| 1       | aad_freiburg | 2.80 (1) | 0.5096 (1) | 0.6059 (4) | 0.6270 (3) | 0.5802 (1) | 0.8778 (5) | 5988.07 (3)  |
| 2       | jrl44        | 3.80 (2) | 0.4856 (2) | 0.6276 (1) | 0.5993 (5) | 0.5292 (3) | 0.8711 (8) | 5986.54 (4)  |
| 3       | tadej        | 4.20 (3) | 0.4309 (9) | 0.6207 (3) | 0.7468 (1) | 0.5549 (2) | 0.8749 (6) | 2727.99 (61) |

# Conclusion

- I believe we shouldn't think in terms of **data scientist hours spent for tuning knobs**, but rather **CPU hours spent for tuning knobs**.
- The human part of the effort is in designing sensible model and model parameter spaces (probably won't be automated soon).

# Conclusion

- I believe we shouldn't think in terms of **data scientist hours spent for tuning knobs**, but rather **CPU hours spent for tuning knobs**.
- The human part of the effort is in designing sensible model and model parameter spaces (probably won't be automated soon).

# Conclusion

- I believe we shouldn't think in terms of **data scientist hours spent for tuning knobs**, but rather **CPU hours spent for tuning knobs**.
- The human part of the effort is in designing sensible model and model parameter spaces (probably won't be automated soon).