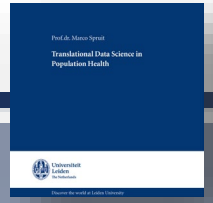
AutoML in Dutch Healthcare

… Towards a Translational Data Science Roadmap

Speaker: *Prof. dr. Marco Spruit (LUMC/LIACS)*

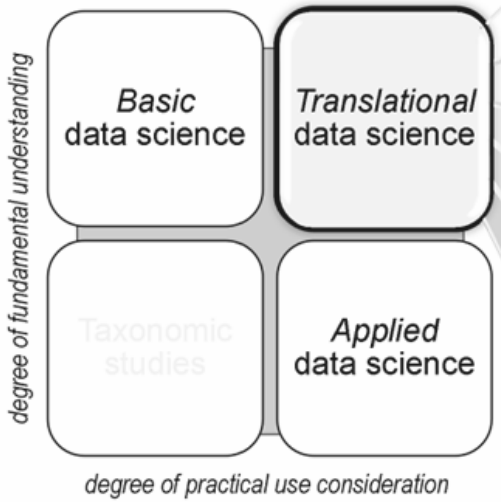
Mission: *"To establish an authoritative national infrastructure for Dutch Natural Language Processing and Machine Learning to democratise Data Science technologies"*

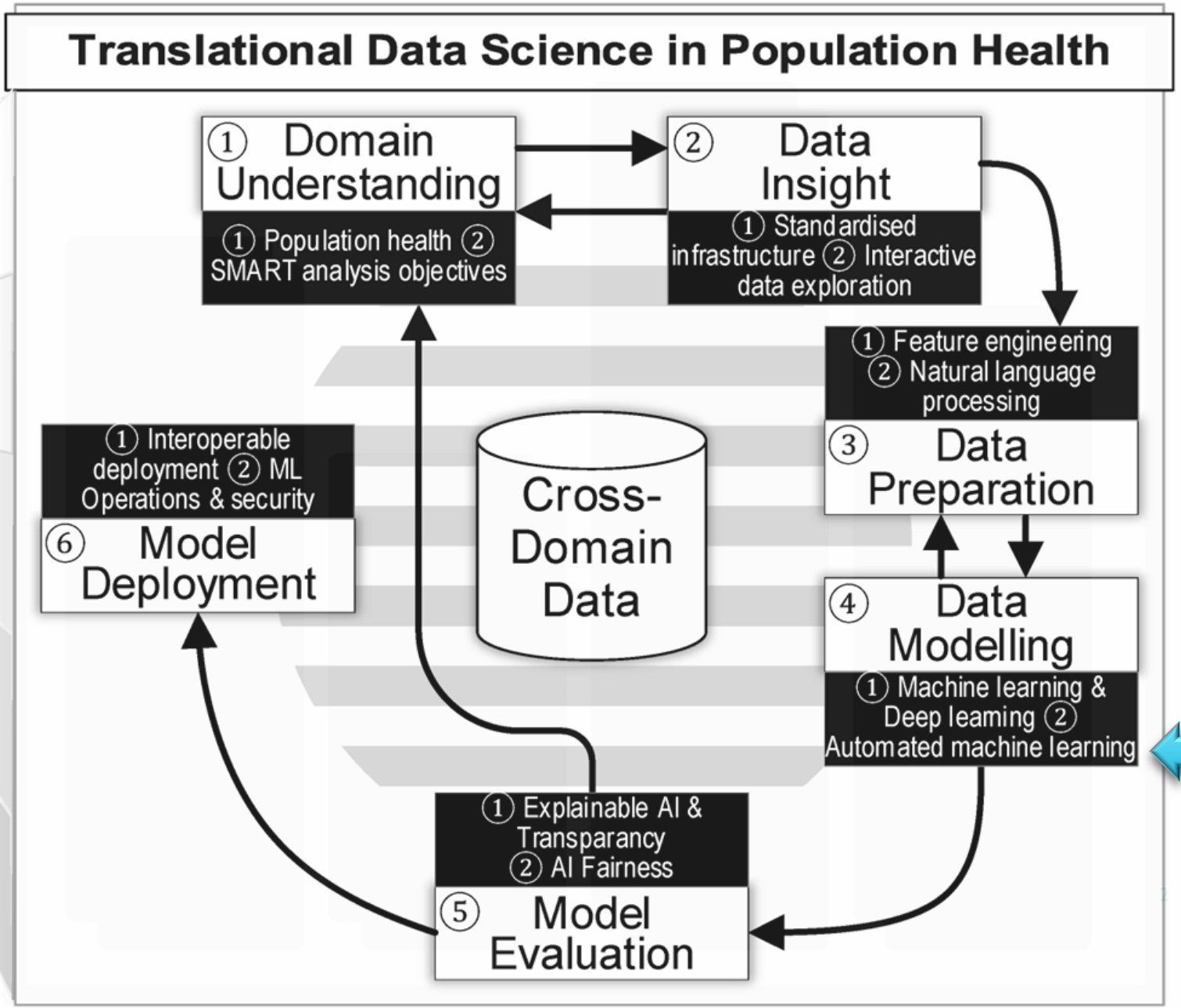
Session: VvE Symposium of SIG Registry-based research
"Let's sail through the numbers"
October 6, UMC Utrecht, Machine Learning session



APRIL FOOLS' DAY

i.e. overview of my TDS research theme





Mission:

"To establish an authoritative national infrastructure for Dutch Natural Language Processing and Machine Learning to democratise Data Science technologies"

Agenda

What is AutoML?

- Top Tools
- Primer

What is AutoML?

What is Machine Learning?

- "A computer program is said to learn from **experience** *E* with respect to some class of **tasks** *T* and **performance** measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

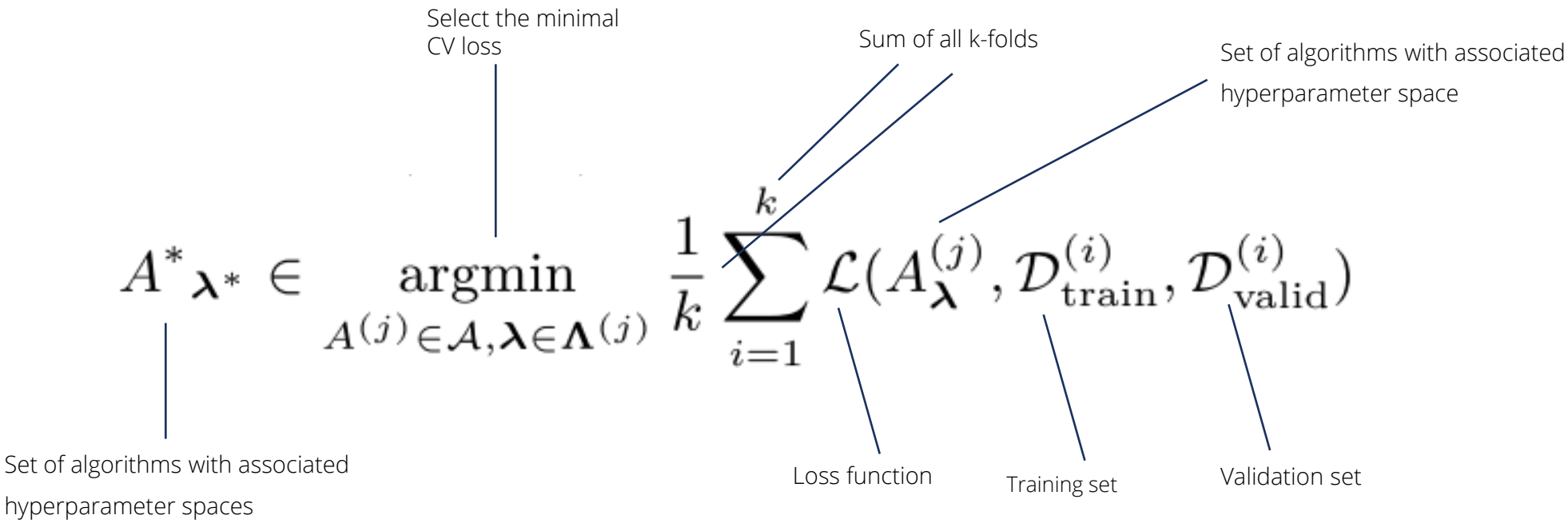
What is **Automated** Machine Learning?

- "AutoML attempts to construct machine learning programs (specified by *E*, *T* and *P*),
  - without human assistance, and
  - within limited computational budgets"

- Mitchell, T. M. (1997:2). Machine learning.-New York, NY, USA: McGraw Hill. Inc. isbn, 70428077.
- Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y. F., Tu, W. W., ... & Yu, Y. (2018). Taking human out of learning applications: A survey on automated machine learning. arXiv preprint arXiv:1810.13306.

THE CASH PROBLEM

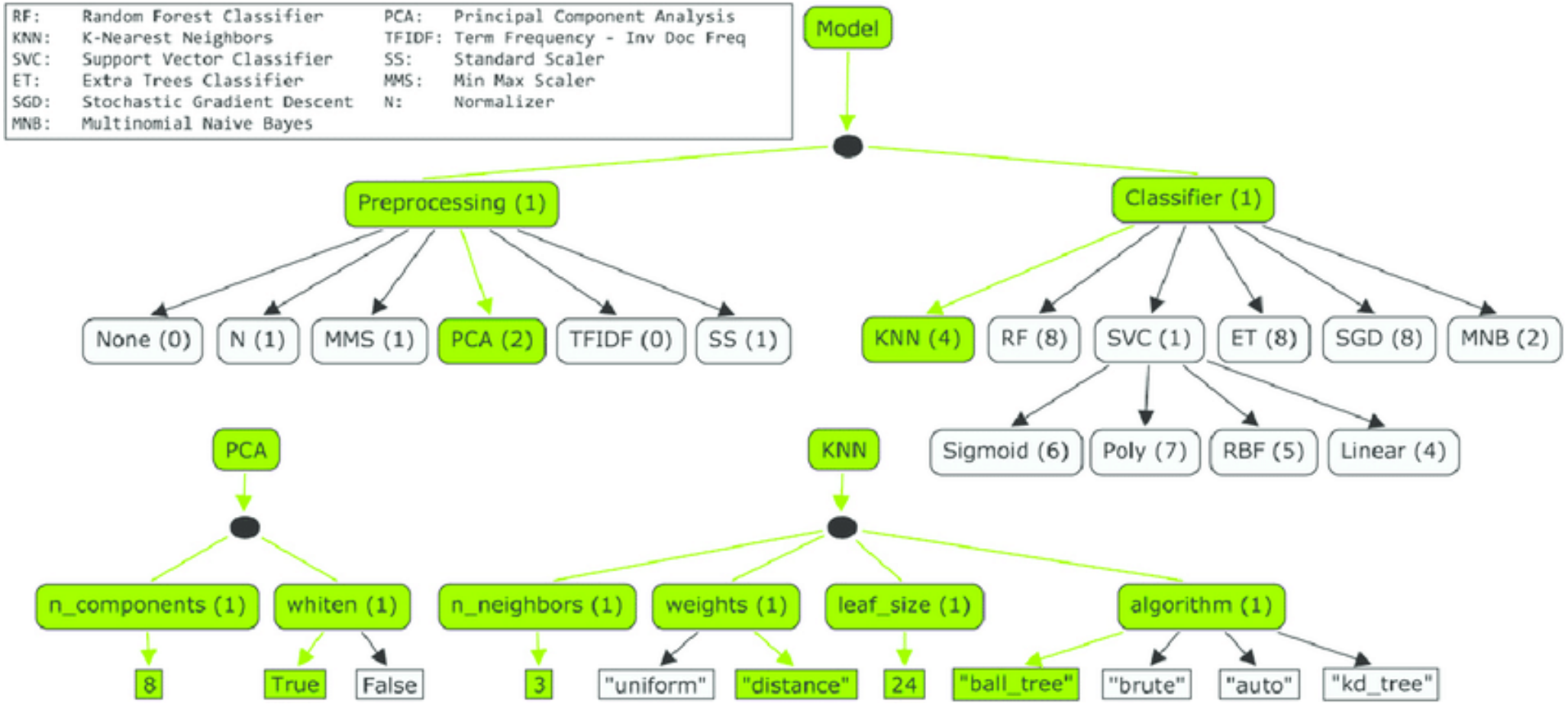
- *a.k.a.* the Combined Algorithm Selection and Hyperparameter optimization problem (CASH)



Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-855).

THE CASH PROBLEM IS A **SEARCH SPACE** PROBLEM

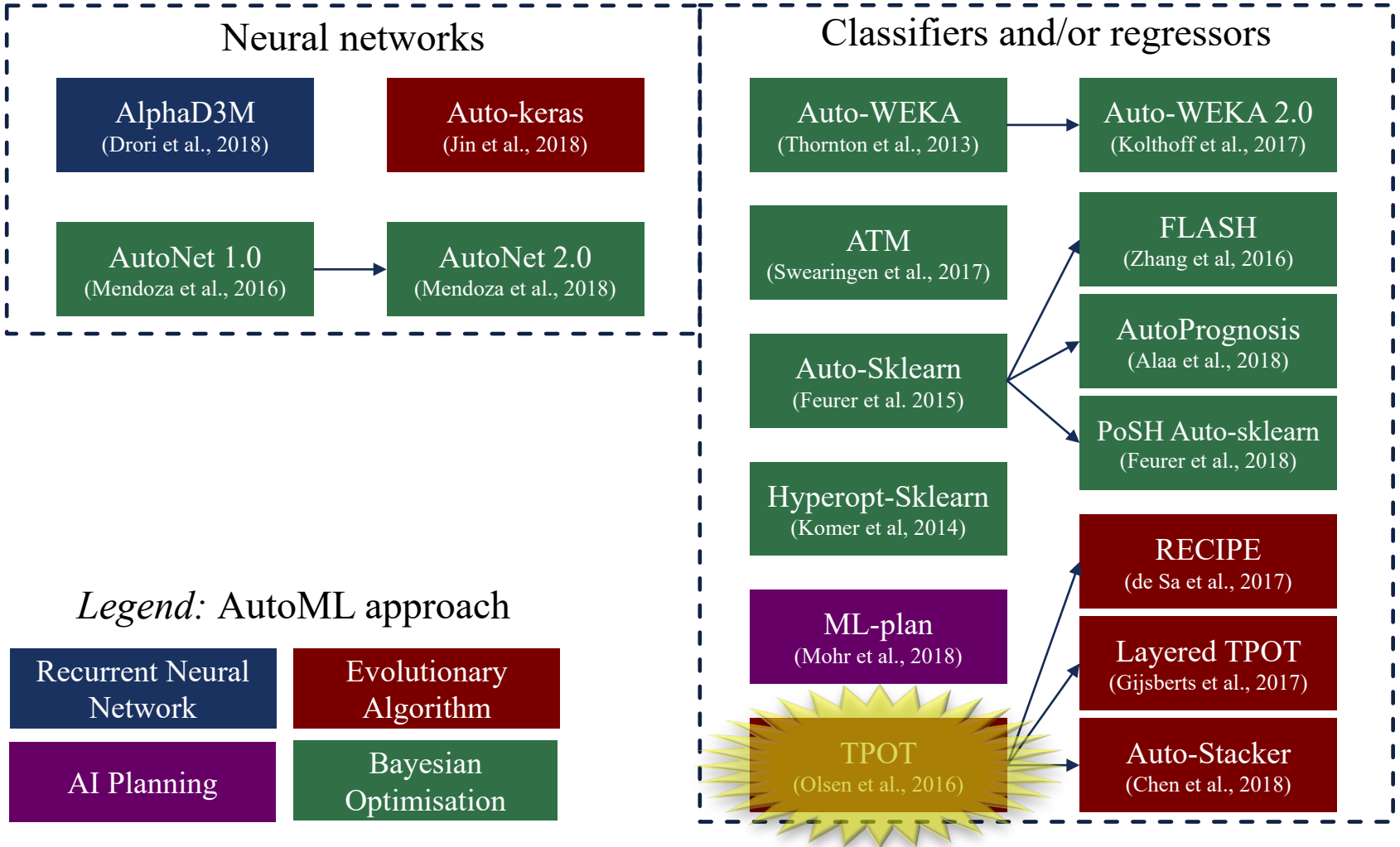
1. Grid & Random search
2. Bayesian Optimization
3. Evolutionary algorithms
4. Reinforcement learning



Bergstra, J., Yamins, D., & Cox, D. D. (2013, June). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference* (Vol. 13, p. 20).

Overview of AutoML methods (2019)

Non-commerical only



Top 10 AutoML Tools in 2022 >> *Automated Machine Learning*

1.
2.
3.
4.
5.

-
-
-

rapidminer

PYCARET

AutoGluon

MLBox,
Machine Learning Box

TPOT

Auto-Sklearn

mljar

AUTO KERAS

H2O.ai

R

AutoWeka
An Automated Data Mining Software Based on Weka

Google Cloud
AutoML Vision

LUDWIG

PyCaret:
Low-Code ML in Python

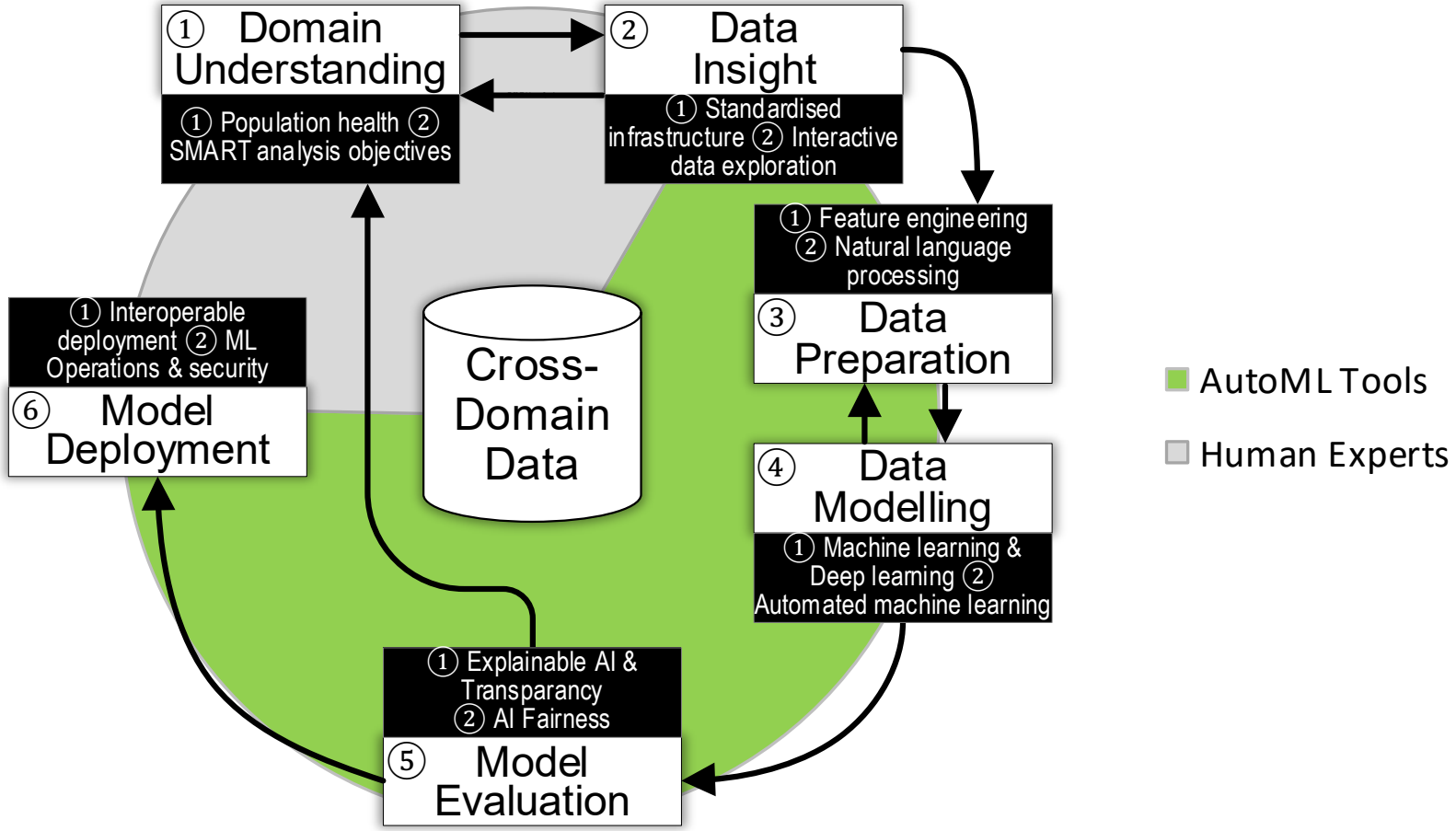
- https://pycaret.org/
- Start Jupyter Notebook…
  - http://localhost:8888/ads-pycaret-book

```
best = compare_models(sort='R2')
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	2342.1429	22959433.7357	4762.0337	0.8351	0.4097	0.2091	0.2360
gbr	Gradient Boosting Regressor	2275.4641	22815428.3119	4750.1089	0.8350	0.3858	0.1874	0.0740
ada	AdaBoost Regressor	3257.2171	23279230.3521	4807.0257	0.8339	0.4770	0.4264	0.0330
lightgbm	Light Gradient Boosting Machine	2491.6919	24030584.9610	4865.2118	0.8272	0.4153	0.2118	0.4180
et	Extra Trees Regressor	2364.4206	25237906.1980	4999.4284	0.8167	0.4283	0.2116	0.2060
catboost	CatBoost Regressor	2530.8745	25732627.8903	5042.4862	0.8134	0.4088	0.2020	1.0060
xgboost	Extreme Gradient Boosting	2931.6919	31946244.2000	5615.7612	0.7678	0.4551	0.2602	0.3410
dt	Decision Tree Regressor	3031.4152	42283353.7664	6468.0098	0.6936	0.5132	0.3181	0.0240
omp	Orthogonal Matching Pursuit	5645.3004	59119654.4986	7679.3606	0.5758	0.6831	0.6880	0.0190
ridge	Ridge Regression	4066.3599	61583179.6000	7714.4257	0.5583	0.4400	0.2707	0.0150
br	Bayesian Ridge	4072.9367	61948316.9816	7735.3075	0.5556	0.4399	0.2705	0.0210
lr	Linear Regression	4081.2541	62419186.8000	7762.5140	0.5521	0.4399	0.2702	0.6810
lar	Least Angle Regression	4081.2284	62418429.6714	7762.4664	0.5521	0.4399	0.2702	0.0210
knn	K Neighbors Regressor	4590.2544	70154126.3724	8271.6145	0.5162	0.5252	0.3131	0.0260
huber	Huber Regressor	4211.3096	80449305.4129	8799.2293	0.4214	0.4535	0.2076	0.0240
par	Passive Aggressive Regressor	5841.9890	94762106.9683	9607.9039	0.2863	0.6406	0.4609	0.0260
en	Elastic Net	8222.6689	160918301.6000	12608.0390	-0.1206	0.9079	0.9707	0.0200
lasso	Lasso Regression	8249.2145	161224220.8000	12619.7366	-0.1227	0.9107	0.9777	0.0150
llar	Lasso Least Angle Regression	8249.2145	161224210.7582	12619.7361	-0.1227	0.9107	0.9777	0.0190

AUTOMATION SCOPE OF AUTOML TOOLS (ESP. PYCARET)

- In the context of Translational Data Science



Agenda: Part II

AutoML in Healthcare?

- Exploratory Case study in Healthcare

Ooms,R., & Spruit,M. (2020). Self-Service Data Science in Healthcare with Automated Machine Learning. *Applied Sciences, 10(*9), Medical Artificial Intelligence, 2992. [https://doi.org/10.3390/app10092992]

AN EXPLORATORY CASE STUDY: AUTOML *IN HEALTHCARE*

Ooms,R., & Spruit,M. (2020). Self-Service Data Science in Healthcare with Automated Machine Learning. *Applied Sciences, 10(9),* Medical Artificial Intelligence, 2992. [https://doi.org/10.3390/app10092992]

- "*How can we support healthcare professionals in their data mining process by applying AutoML?*"
  - Investigate by employing from the OpenML-CC18 open-source benchmark suite of Gijsbers et al. (2019),
  - Include all medical datasets suited for binary classification problems: 4

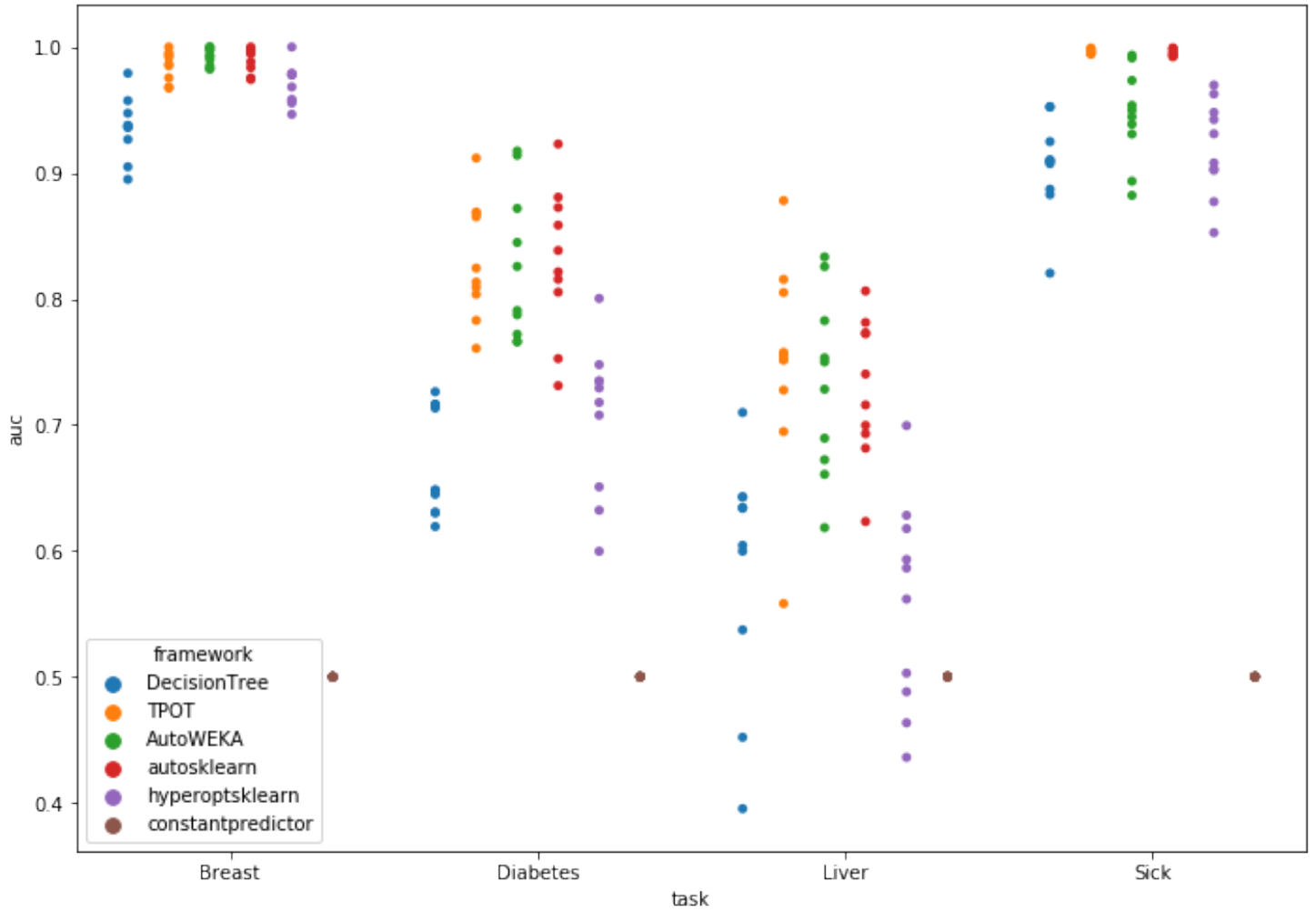
Dataset	Data points	Missing data	Predictive features	Class variable
Breast cancer	699	-	9	458/241
Diabetes	768	-	8	500/268
Indian Liver Patients	583	-	10	416/167
Sick	3772	6064	29	3541/231

- All AutoML methods receive 1 hour in a 10-fold cross-validation set-up to create the best pipeline on these datasets

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An open source AutoML benchmark. *arXiv preprint arXiv:1907.00909.*

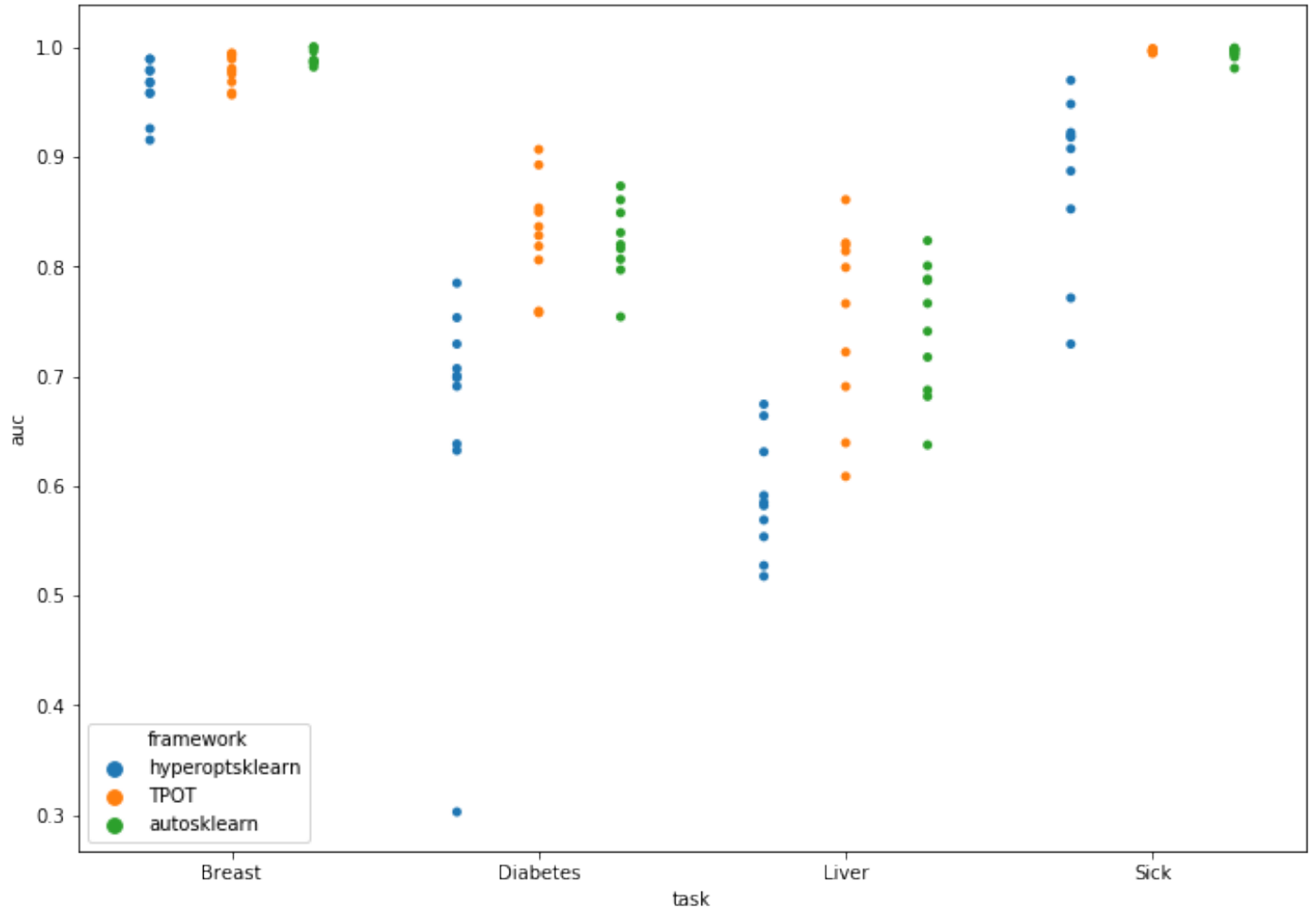
BENCHMARKING WITH 1 HOUR BUDGET

- Decision tree and constant predictor as baseline
- Significant differences per dataset ($p < 0.001$) for all methods
- Hyperopt performs worst
- **TPOT and Auto-Sklearn perform best**



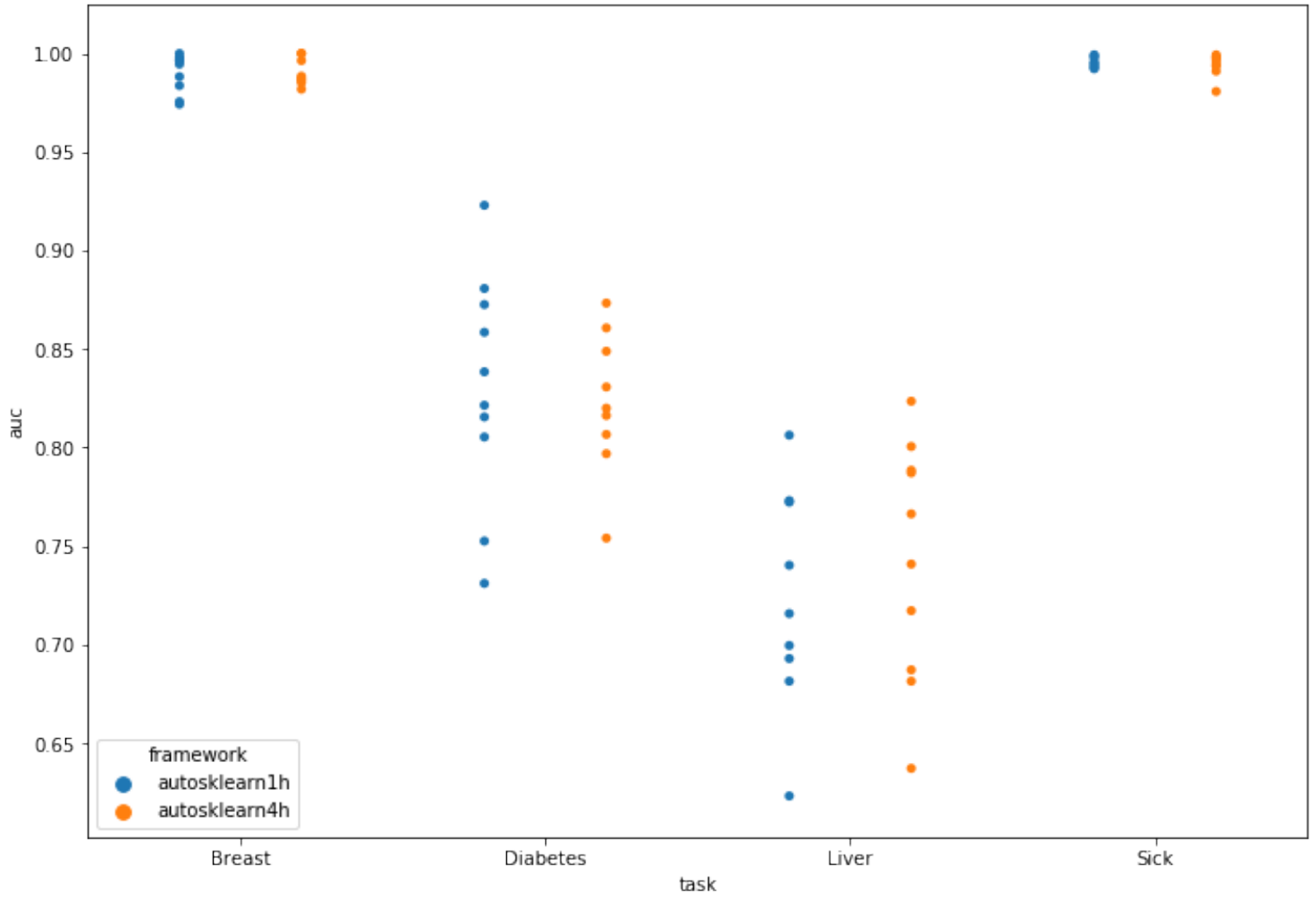
BENCHMARKING WITH **4 HOUR BUDGET**

- Significant differences per dataset ($p < 0.001$) for all methods
- **TPOT and Auto-Sklearn perform best (again)**

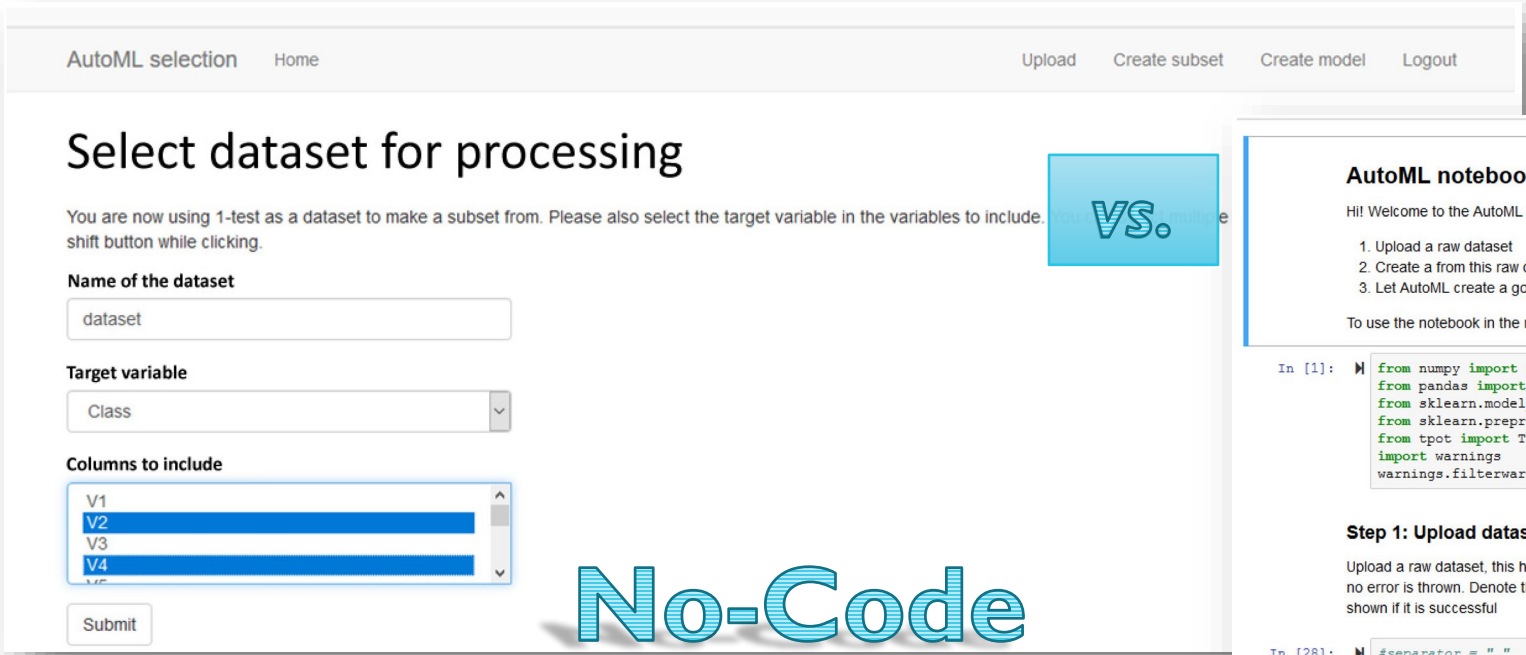


BENCHMARKING AUTO-SKLEARN **BUDGETS**

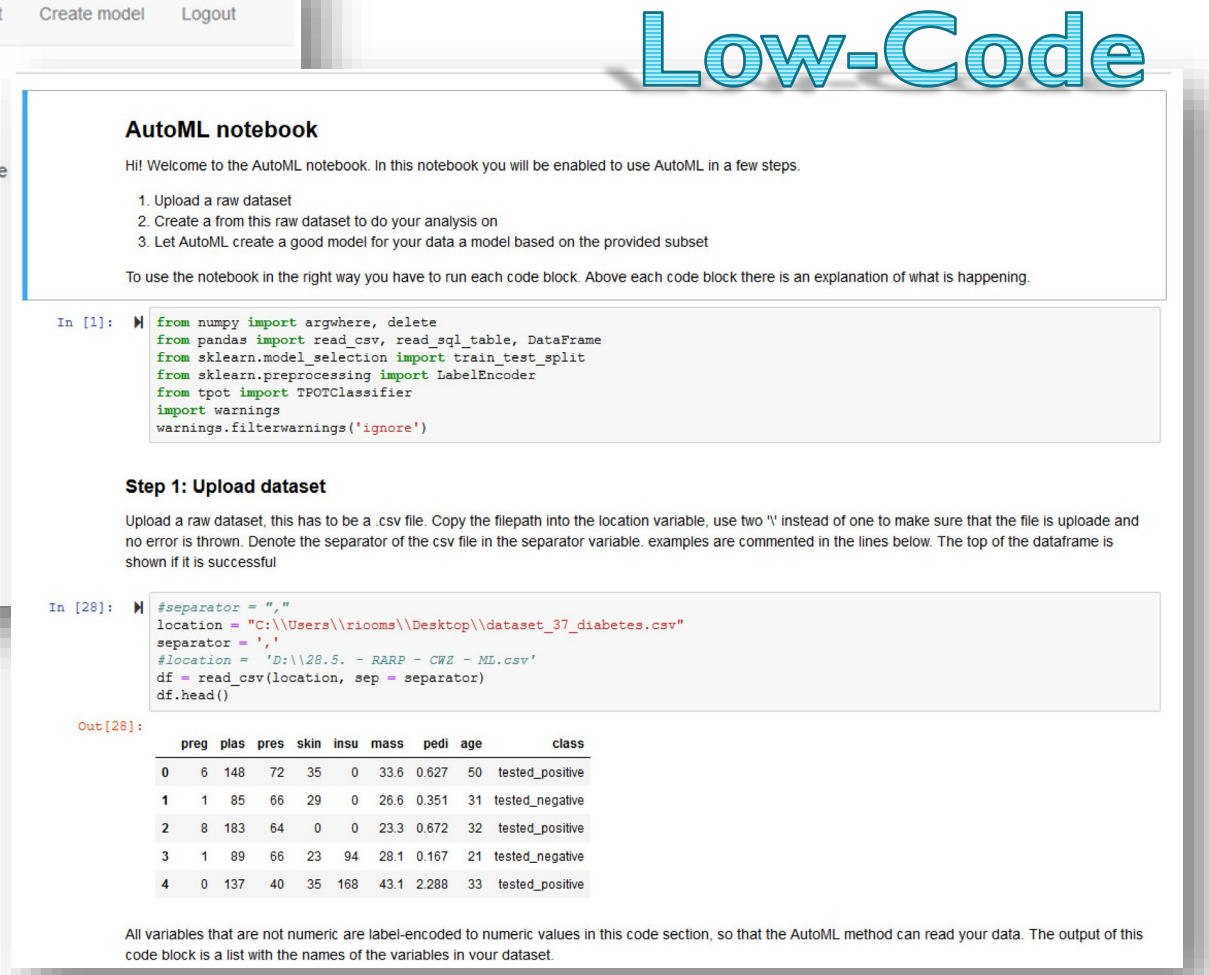
- Increase in time budget does **not** improve performance!



Evaluation: Webapp vs Notebook deployment



vs.



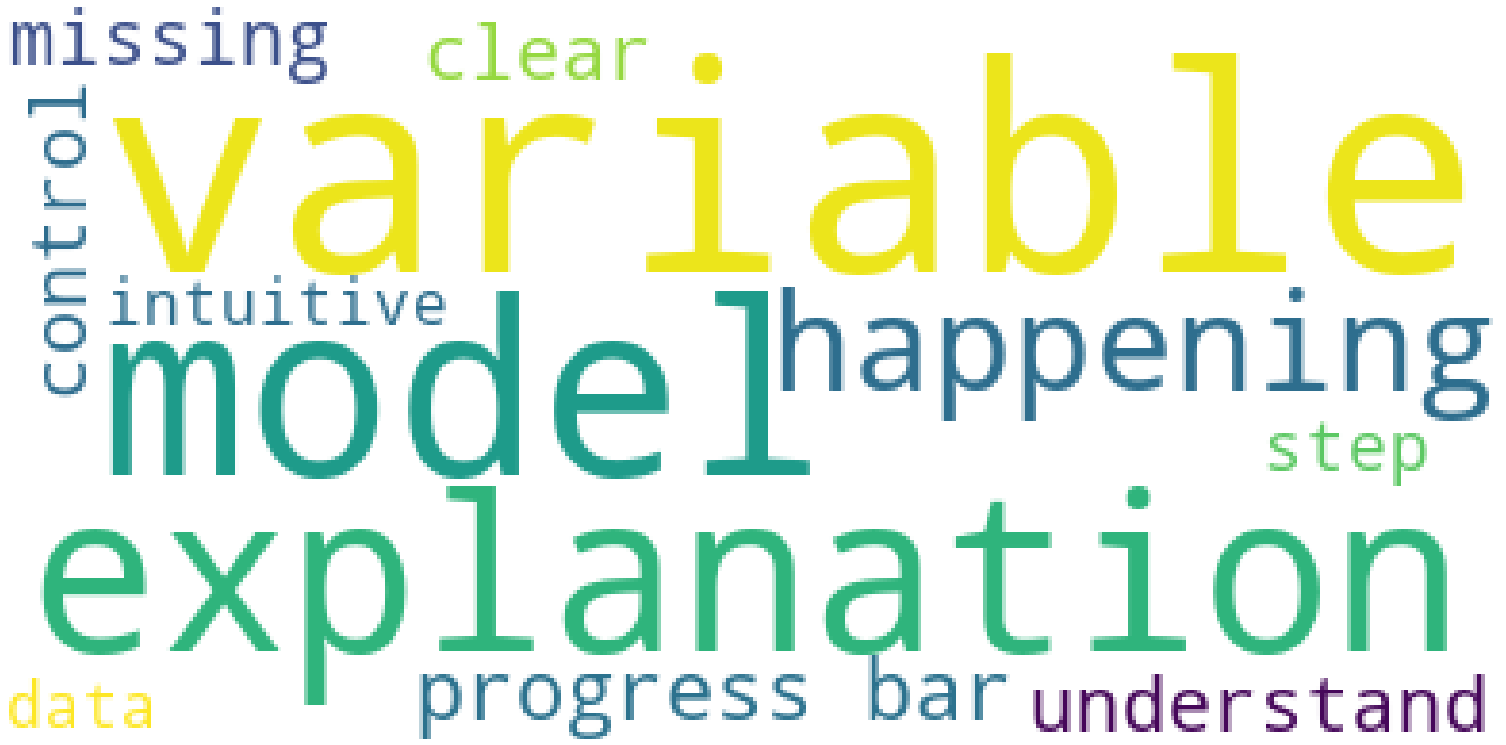
Artefact **A**: Graphical User interface ^^
Flask web application

Artefact **B**: Interactive code/text interface
Jupiter notebook >>

Ooms,R. & **Spruit,M.** (2020). Self-Service Data Science in Healthcare with Automated Machine Learning. *Applied Sciences*, *10*(9), Medical Artificial Intelligence, 2992. [JIF: 2.679] [pdf] [online]

Evaluation: Requirements of 5 senior Researcher-physicians

- The 5 interviewees showed a preference for ***explainability*** of both model construction and model explanation



Evaluation: Empirical Findings

- A hybrid version of the two artefacts would be the perfect fit for the subjects…
- Artefact A (webapp) is preferred for 'basic' operations and overview
- Artefact B (notebook) is preferred for model construction

Code-based interface	Graphical user interface
More control	Less control
More error prone	Less errors

Category	*Preference*	*Score*
User interaction		
Upload dataset	A	4/5
Create a subset	B	3/5
Workflow	B	4/5
Workflow explanation	B	4/5
Model construction		
Progress reporting	B	4/5
Model construction	B	5/5
Model explanation		
Compare results	A	4/5
Explanation missing data	B	4/5
Readability	A	4/5

Some Interview Quotes

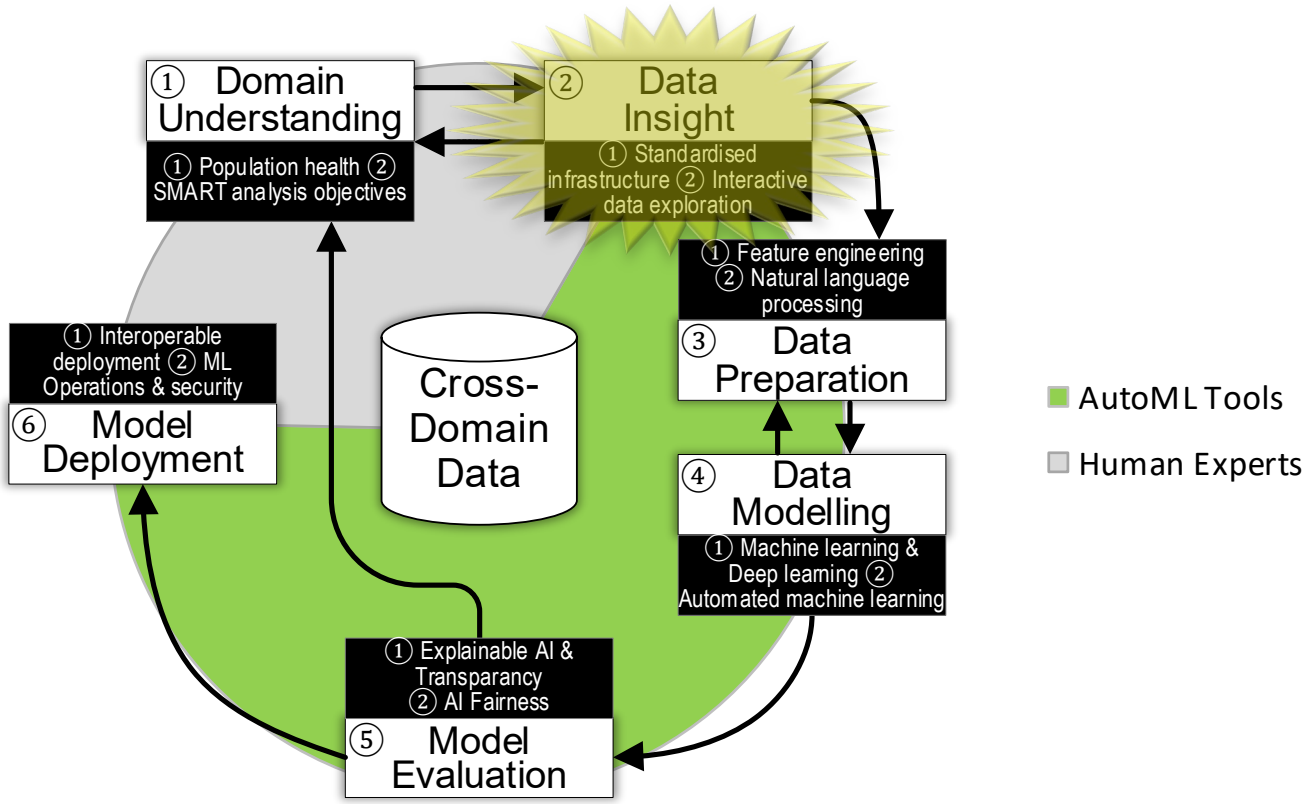
- *"The Webapp (A) overview [of comparing results] is clearer"*
- *"The Notebook (B) provides me with more insight in what I am doing"*
- *"I do not understand all the code [in B].*

  However, being able to see the code makes it feel like I am more in control"

- *"It [TPOT output] does not show an answer to my question. I do not consider this a model."*
- *"Difficult to understand for me, I want to know the coefficients of the variables"*
- *"Let me input my time constraint options. - I want more information and choice on the type of output."*

Conclusion

- AutoML in its current state is *unable* to support researcher-physicians in their knowledge discovery process because modelling decisions and variable importance are not shared.
- However, it *can* be used in the data understanding phase of their knowledge discovery process.



THANKS 👍

- Contact? m.r.spruit@lumc.nl

??? But.... Would you...???