

# 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



1993



1995



1997



2003

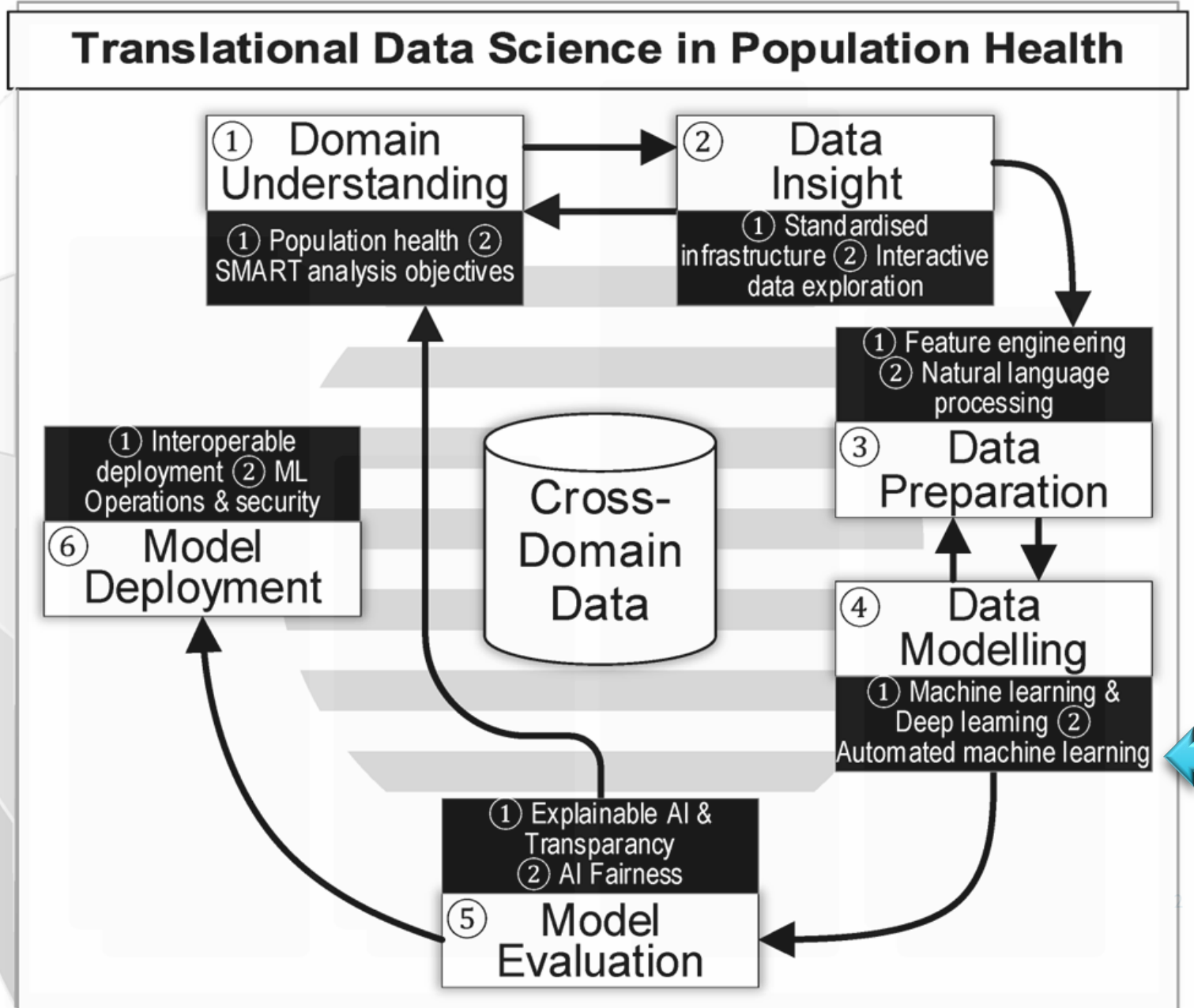
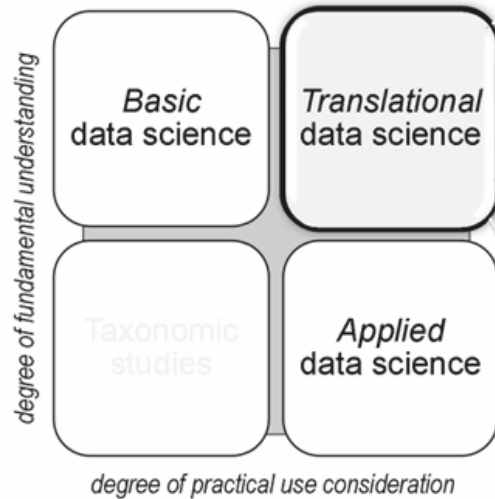


2007



# 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 AUTO ML?

## 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  $E$ .”
- Mitchell, T. M. (1997:2). Machine learning.-New York, NY, USA: McGraw Hill. Inc. isbn, 70428077.



## 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”
- 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)

The diagram illustrates the CASH problem equation with several annotations:

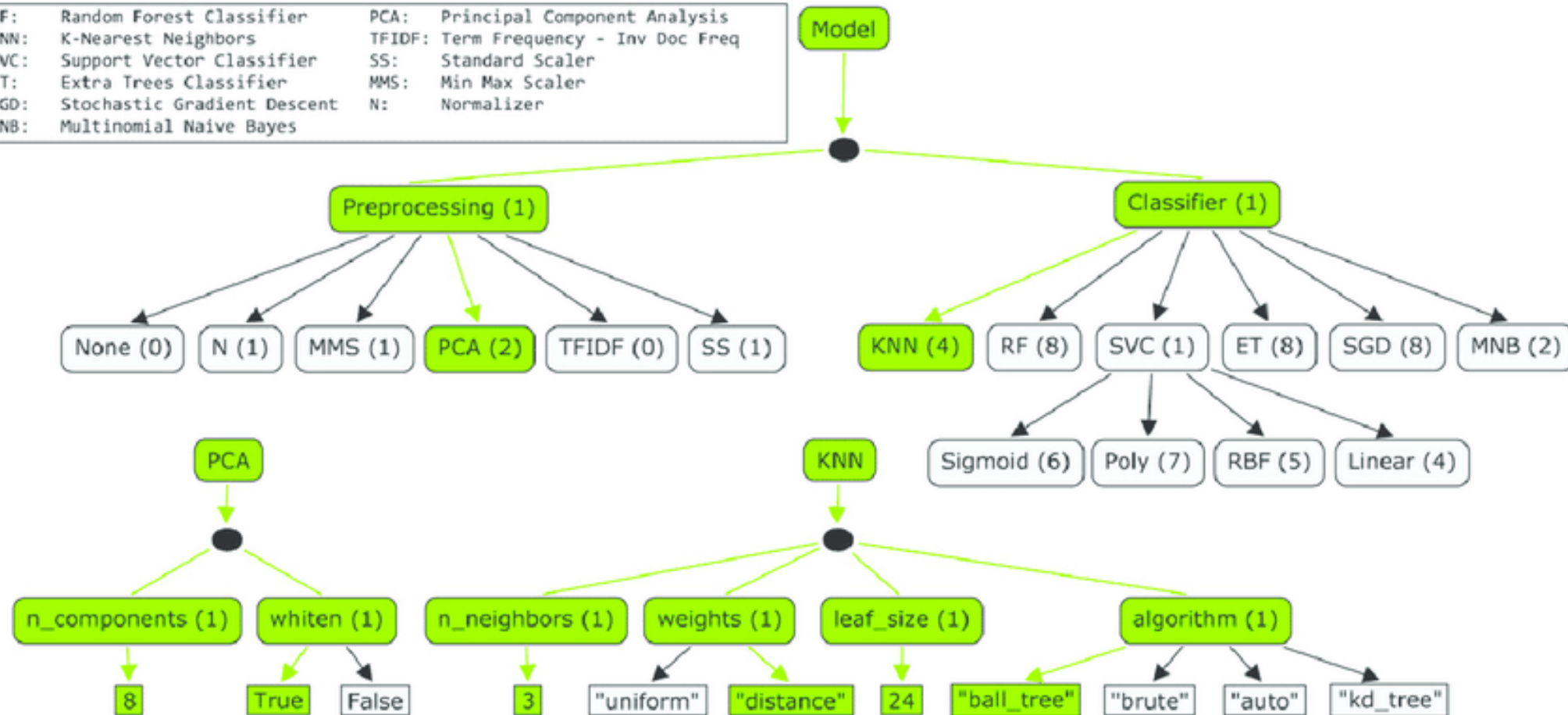
- Set of algorithms with associated hyperparameter spaces**: Points to the term  $A^* \lambda^* \in$ .
- Select the minimal CV loss**: Points to the  $\operatorname{argmin}$  operator.
- Sum of all k-folds**: Points to the summation symbol  $\sum_{i=1}^k$ .
- Loss function**: Points to the  $\mathcal{L}$  symbol.
- Training set**: Points to the  $\mathcal{D}_{\text{train}}^{(i)}$  term.
- Validation set**: Points to the  $\mathcal{D}_{\text{valid}}^{(i)}$  term.
- Set of algorithms with associated hyperparameter space**: Points to the  $A_{\lambda}^{(j)}$  term.

$$A^* \lambda^* \in \operatorname{argmin}_{A^{(j)} \in \mathcal{A}, \lambda \in \Lambda^{(j)}} \frac{1}{k} \sum_{i=1}^k \mathcal{L}(A_{\lambda}^{(j)}, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)})$$

# THE CASH PROBLEM IS A SEARCH SPACE PROBLEM

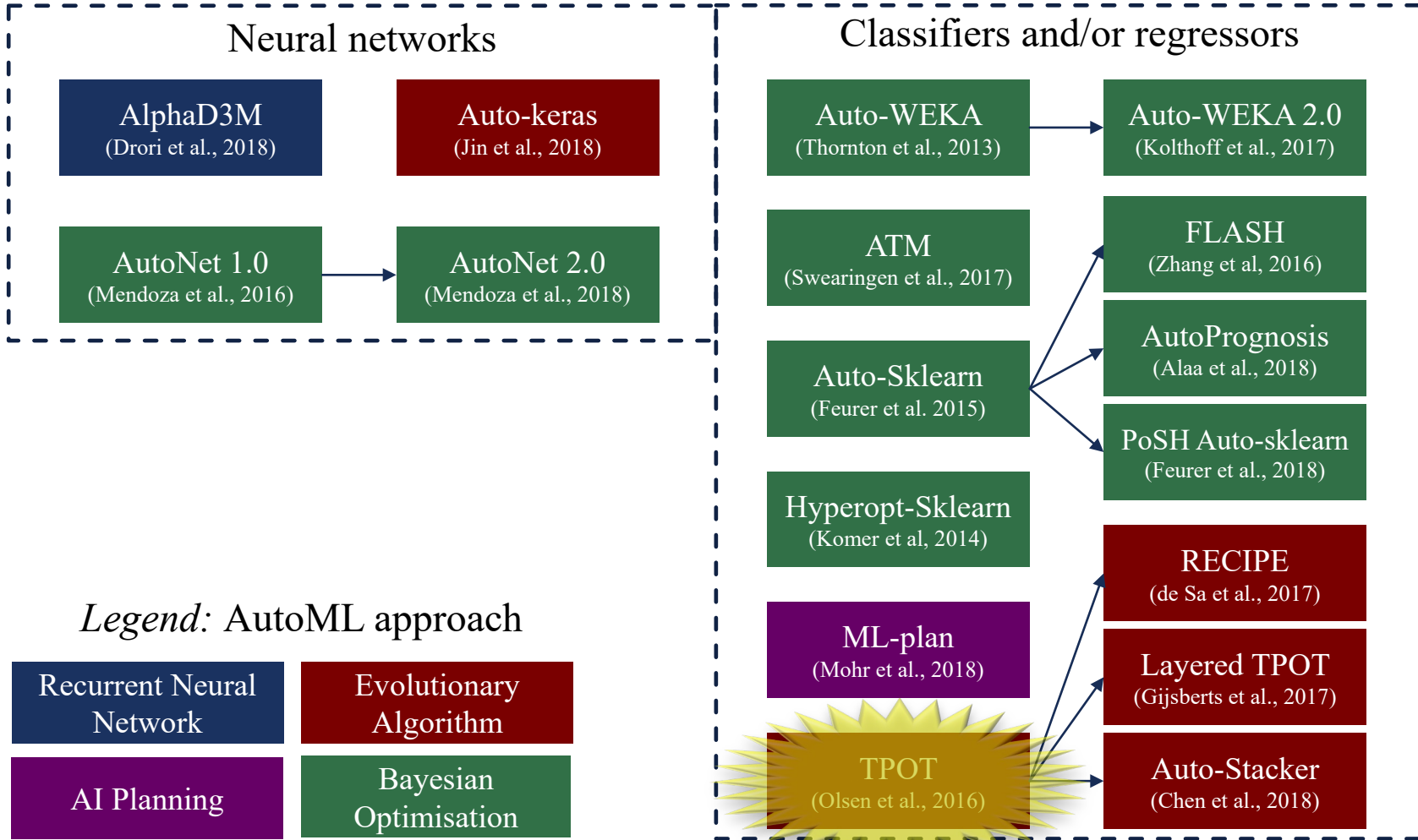
1. Grid & Random search
2. Bayesian Optimization
3. Evolutionary algorithms
4. Reinforcement learning

RF:	Random Forest Classifier	PCA:	Principal Component Analysis
KNN:	K-Nearest Neighbors	TFIDF:	Term Frequency - Inv Doc Freq
SVC:	Support Vector Classifier	SS:	Standard Scaler
ET:	Extra Trees Classifier	MMS:	Min Max Scaler
SGD:	Stochastic Gradient Descent	N:	Normalizer
MNB:	Multinomial Naive Bayes		



Non-commercial only

# OVERVIEW OF AUTO ML METHODS (2019)



# TOP 10 AUTOML TOOLS IN 2022 >> AUTOMATED MACHINE LEARNING

1.



2.



3.



**MLBox,  
Machine Learning Box**

4.



5.

**Auto-Sklearn**

Disclaimer



**AUTO KERAS**



■

■

■



**Google Cloud  
AutoML Vision**



# PYCARET: LOW-CODE ML IN PYTHON

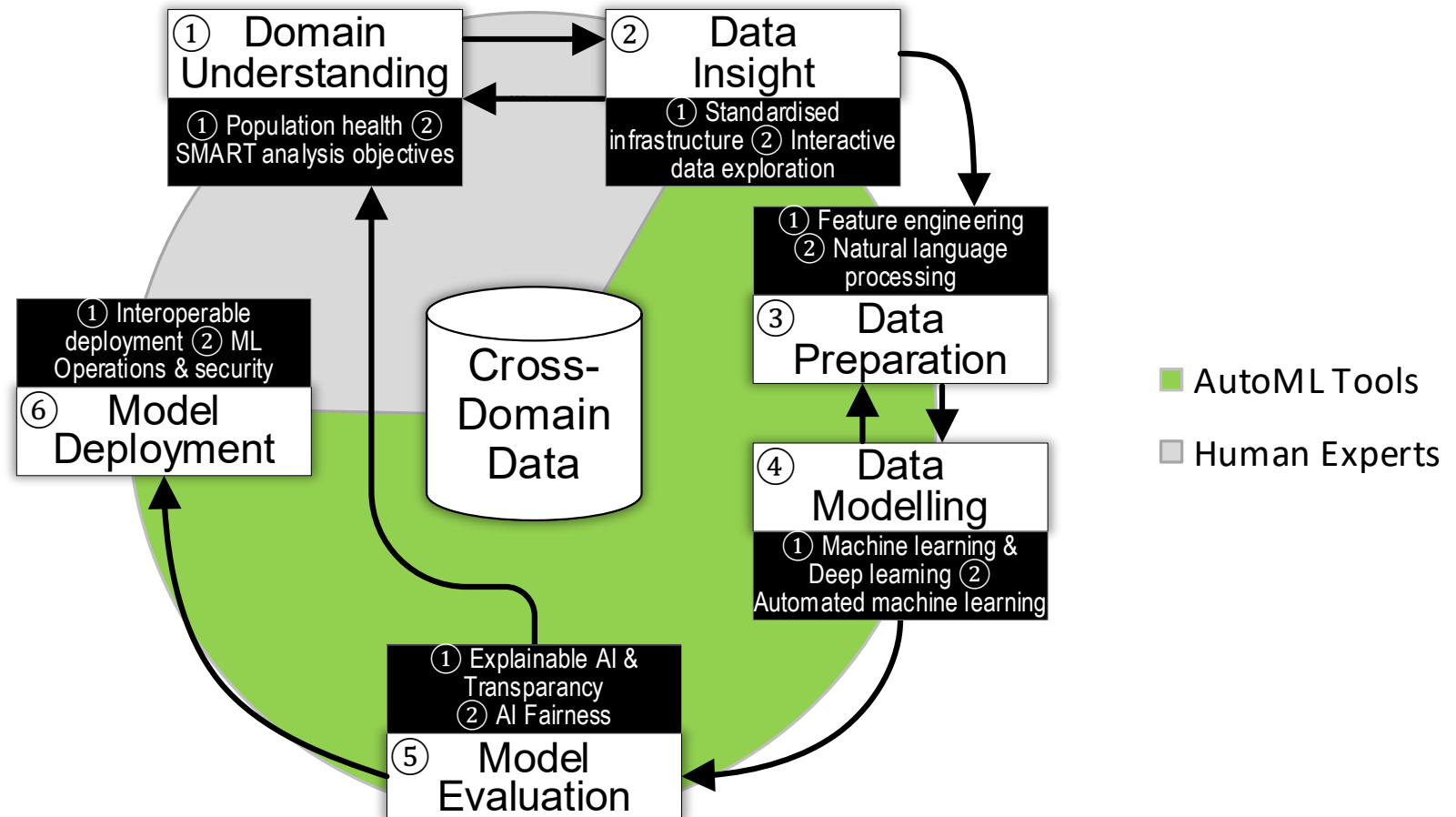
- <https://pycaret.org/>
- Start Jupyter Notebook...
  - <http://localhost:8888/ads-pycaret-book>

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

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

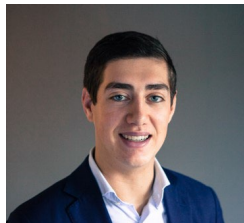
- In the context of Translational Data Science



# 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>]



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: AUTO ML IN HEALTHCARE

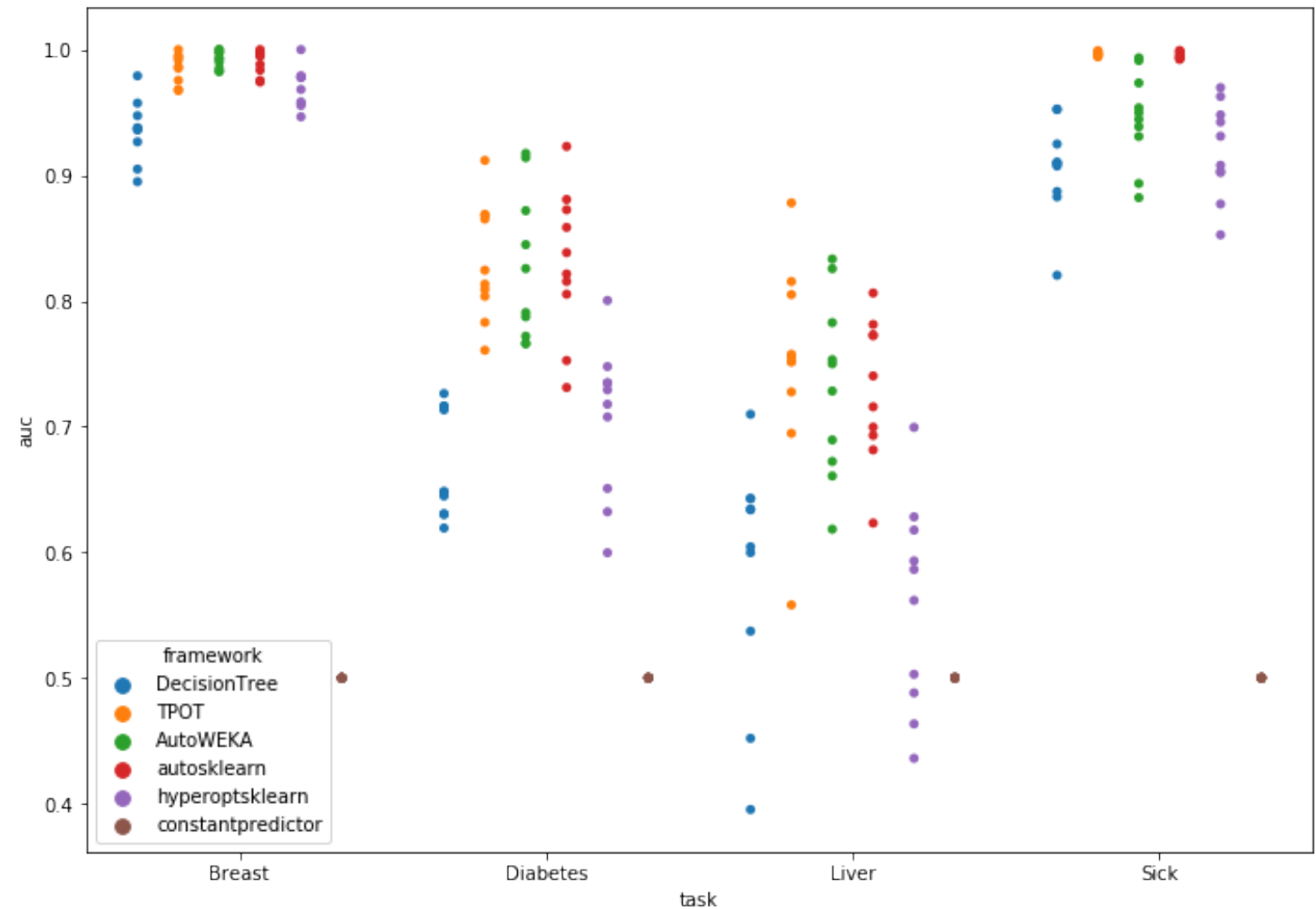
- “How can we support healthcare professionals in their data mining process by applying AutoML?”
  - Investigate by employing from the OpenML-CC18 open-source benchmark suite of Gijssbers et al. (2019),
  - Include all medical datasets suited for binary classification problems: 4

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

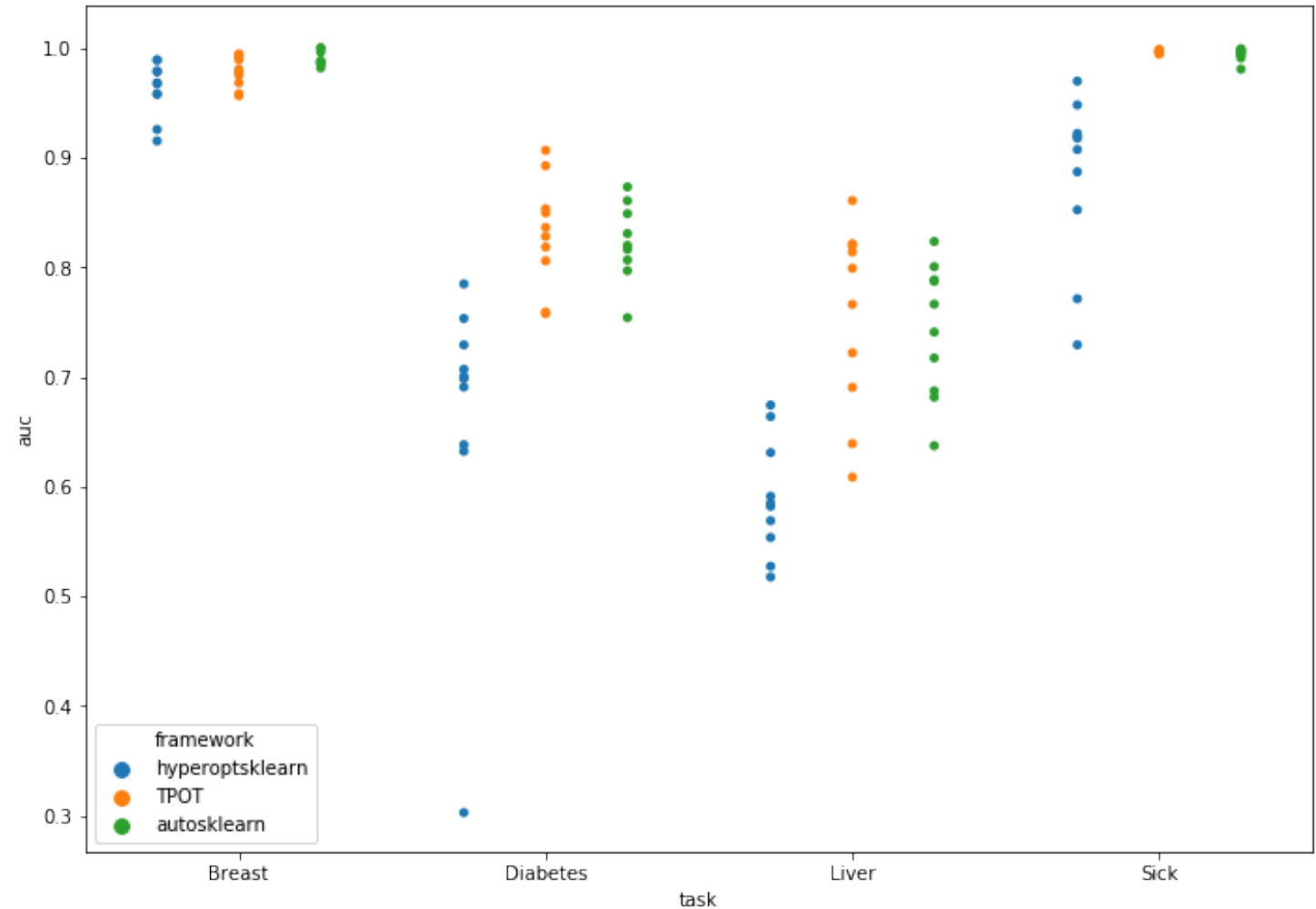
# BENCHMARKING WITH 1 HOUR BUDGET

- 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



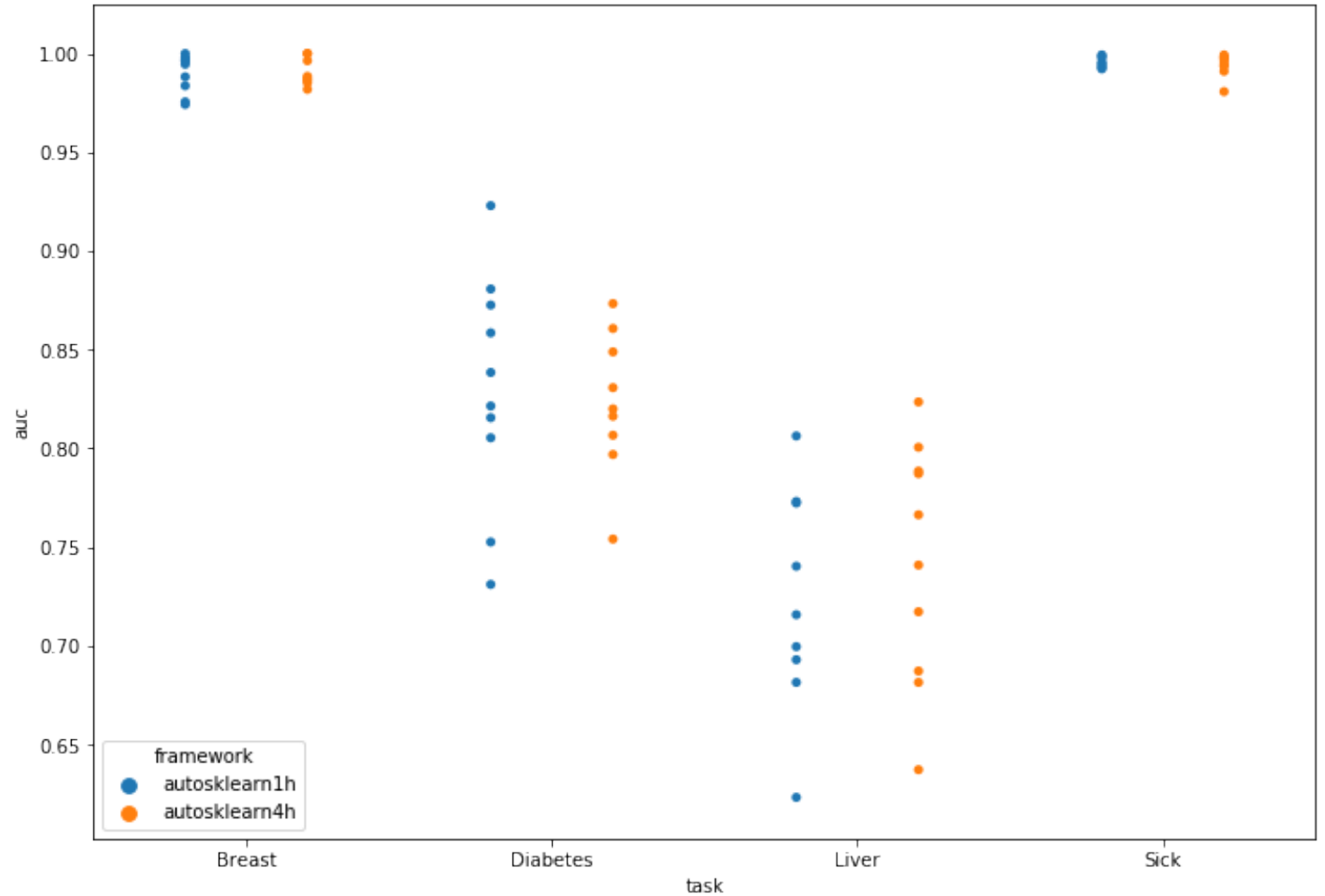
# BENCHMARKING WITH 4 HOUR BUDGET

- 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

AutoML selection Home Upload Create subset Create model Logout

## Select dataset for processing

You are now using 1-test as a dataset to make a subset from. Please also select the target variable in the variables to include. shift button while clicking.

**Name of the dataset**  
dataset

**Target variable**  
Class

**Columns to include**  
V1  
V2  
V3  
V4  
V5

Submit

VS.

No-Code

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

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

Low-Code

## AutoML notebook

Hi! Welcome to the AutoML notebook. In this notebook you will be enabled to use AutoML in a few steps.

1. Upload a raw dataset
2. Create a subset from this raw dataset to do your analysis on
3. Let AutoML create a good model for your data a model based on the provided subset

To use the notebook in the right way you have to run each code block. Above each code block there is an explanation of what is happening.

```
In [1]: from numpy import argwhere, delete
from pandas import read_csv, read_sql_table, DataFrame
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tpot import TPOTClassifier
import warnings
warnings.filterwarnings('ignore')
```

### Step 1: Upload dataset

Upload a raw dataset, this has to be a .csv file. Copy the filepath into the location variable, use two '\ ' instead of one to make sure that the file is uploaded and no error is thrown. Denote the separator of the csv file in the separator variable. examples are commented in the lines below. The top of the dataframe is shown if it is successful

```
In [28]: #separator = ","
location = "C:\\Users\\zirooms\\Desktop\\dataset_37_diabetes.csv"
separator = ','
#location = 'D:\\28.5. - RARP - CWS - ML.csv'
df = read_csv(location, sep = separator)
df.head()
```

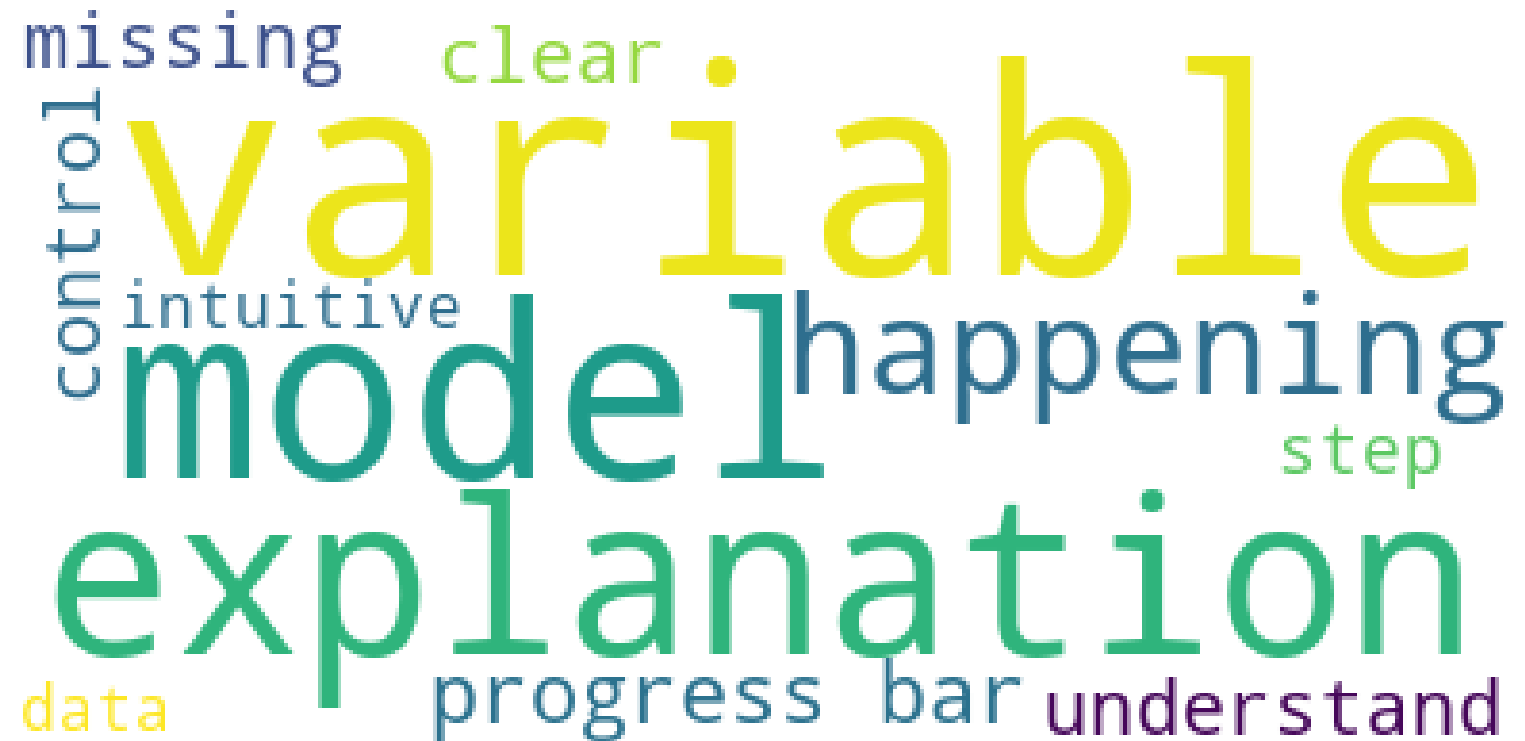
```
Out[28]:
```

	preg	plas	pres	skin	insu	mass	pedi	age	class
0	6	148	72	35	0	33.6	0.627	50	tested_positive
1	1	85	66	29	0	26.6	0.351	31	tested_negative
2	8	183	64	0	0	23.3	0.672	32	tested_positive
3	1	89	66	23	94	28.1	0.167	21	tested_negative
4	0	137	40	35	168	43.1	2.288	33	tested_positive

All variables that are not numeric are label-encoded to numeric values in this code section, so that the AutoML method can read your data. The output of this code block is a list with the names of the variables in your dataset.

# 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

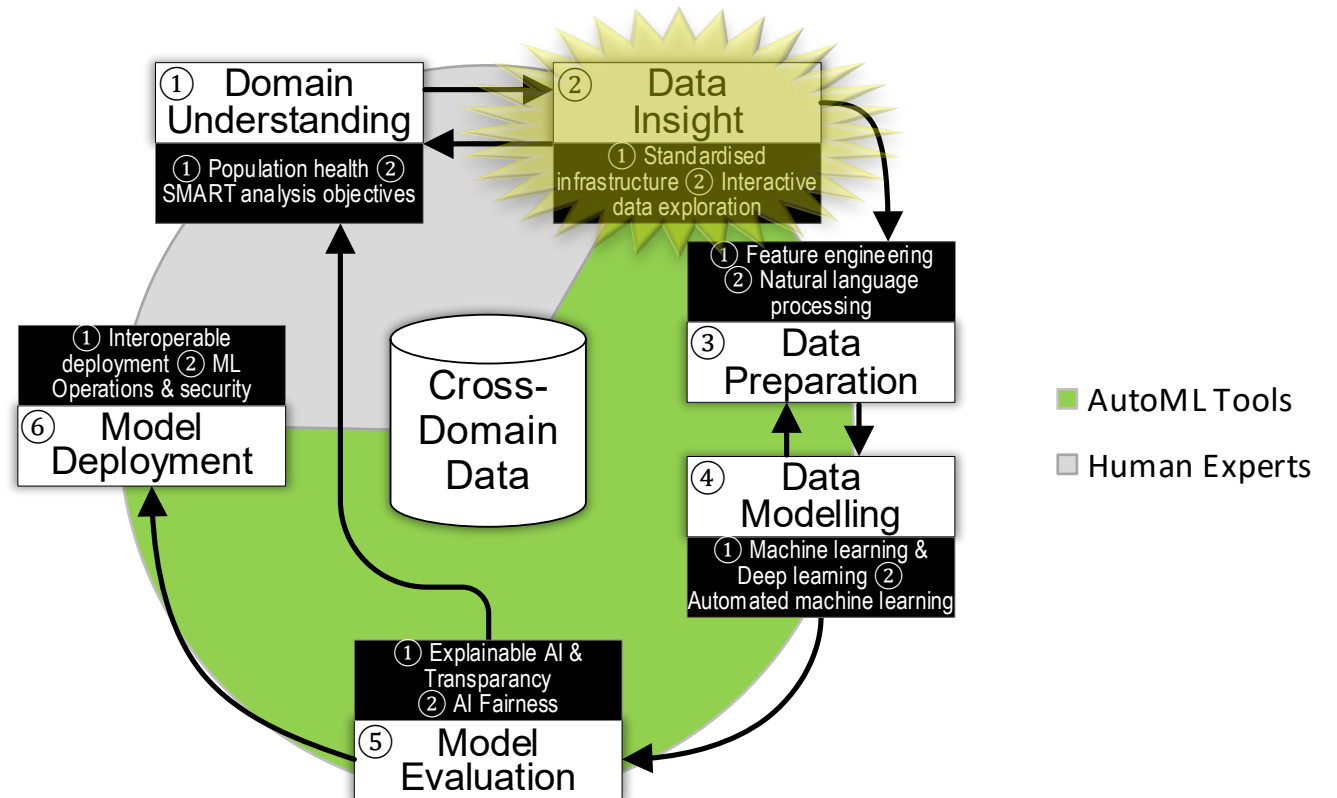
Category	Preference	Score
<b>User interaction</b>		
Upload dataset	A	4/5
Create a subset	B	3/5
Workflow	B	4/5
Workflow explanation	B	4/5
<b>Model construction</b>		
Progress reporting	B	4/5
Model construction	B	5/5
<b>Model explanation</b>		
Compare results	A	4/5
Explanation missing data	B	4/5
Readability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

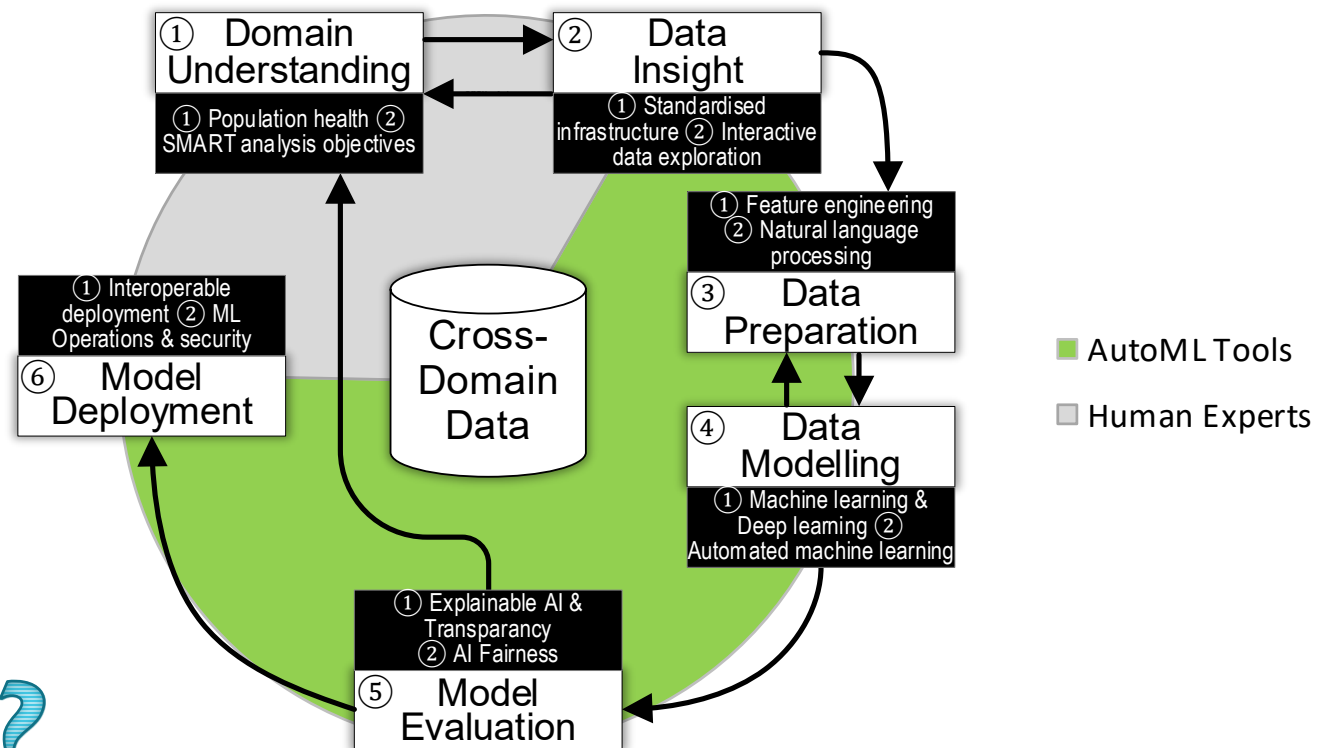
# CONCLUSION

- 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](mailto:m.r.spruit@lumc.nl)



???

But...

Would you...???