Elements of Machine Learning & Data Science

Winter semester 2023/24

Automated Machine Learning (4)

Prof. Holger Hoos

Last week:

- How to optimise hyperparameters of a ML model / training procedure?

  -> hyperparameter optimisation (HPO)

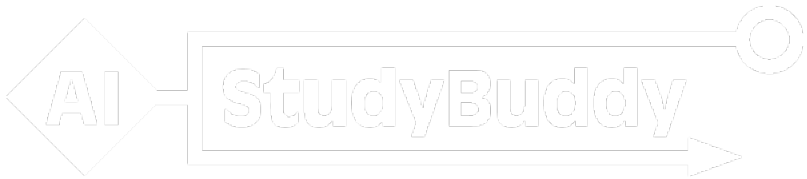
- How to optimise neural network architectures? (NAS)

  -> neural architecture search (NAS)

TPS Exercise

Question:

What are the key differences between HPO and NAS?



TPS Exercise

Question:

What are the key differences between HPO and NAS?

- Numerical vs categorical parameters + parameter dependencies
- Open-endedness of search space (NAS)
- Limitation to neural networks (NAS)

Preparation for today:

Read the following research paper:

Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, Christopher K. I. Williams: Automating Data Science. Communications of the ACM, March 2022, Vol. 65 No. 3, Pages 76-87

(The paper is available online at https://dl.acm.org/doi/pdf/10.1145/3495256)

Focus on the following questions (which will be further explored in TPS exercises in class):

(1) What are the three forms of automation relevant in the context of data science, and why do they all matter?

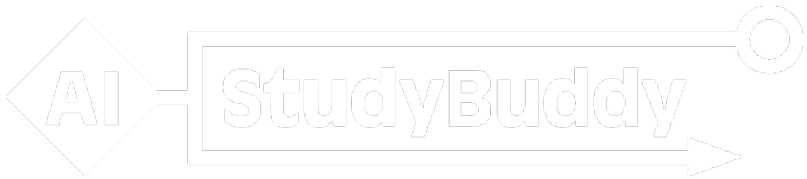
(2) What are the main challenges in the area of data engineering?

(3) Which of the four quadrants of data science are the most challenging to automate?

Bring your answers to these questions (which can be in the form of bullet points) to class; they will be the basis for TSP exercises).

Today:

- How to automatically construct / optimise ML pipelines?
- More broadly, how to automate data science?



TPS Exercise

End-to-end learning using neural networks is widely regarded the cutting edge of machine learning.

Question:

Under what circumstance might it be wise / necessary to choose a different approach / model?

TPS Exercise

End-to-end learning using neural networks is widely regarded the cutting edge of machine learning.

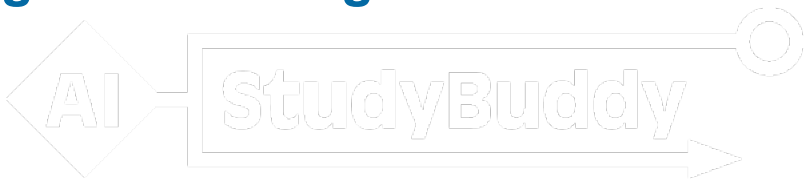
Question:

Under what circumstance might it be wise / necessary to choose a different approach / model?

- Limited amount of data
- Model needs to be explainable
- Limited amount of computing resources for training / deployment (e.g., edge devices)
- Temporal data (currently needs dedicated pre-processing)

How to automatically construct / optimise ML pipelines?

- Specify possible components (e.g., preprocessing / training procedures) and their hyperparameters
  + ways to combine them
- Use a general-purpose algorithm configurator to instantiate these design choices



-> Auto-WEKA (Thornton et al. 2013),
Auto-sklearn (Feurer et al. 2015),

…

Algorithm selection:

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

Combined algorithm selection and hyperparameter optimisation (CASH):

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

(see Thornton et al. 2013 for details – but beware, *k* is overloaded)

Key idea for solving CASH using general-purpose algorithm configurators:

Define additional categorical hyperparameter to select algorithm $A^{(j)}$

TPS Exercise

Question:

In which respect is data science more than just machine learning?

Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...

Model Building:
algorithm selection,
parameter optimization,
performance evaluation,
model selection,
...

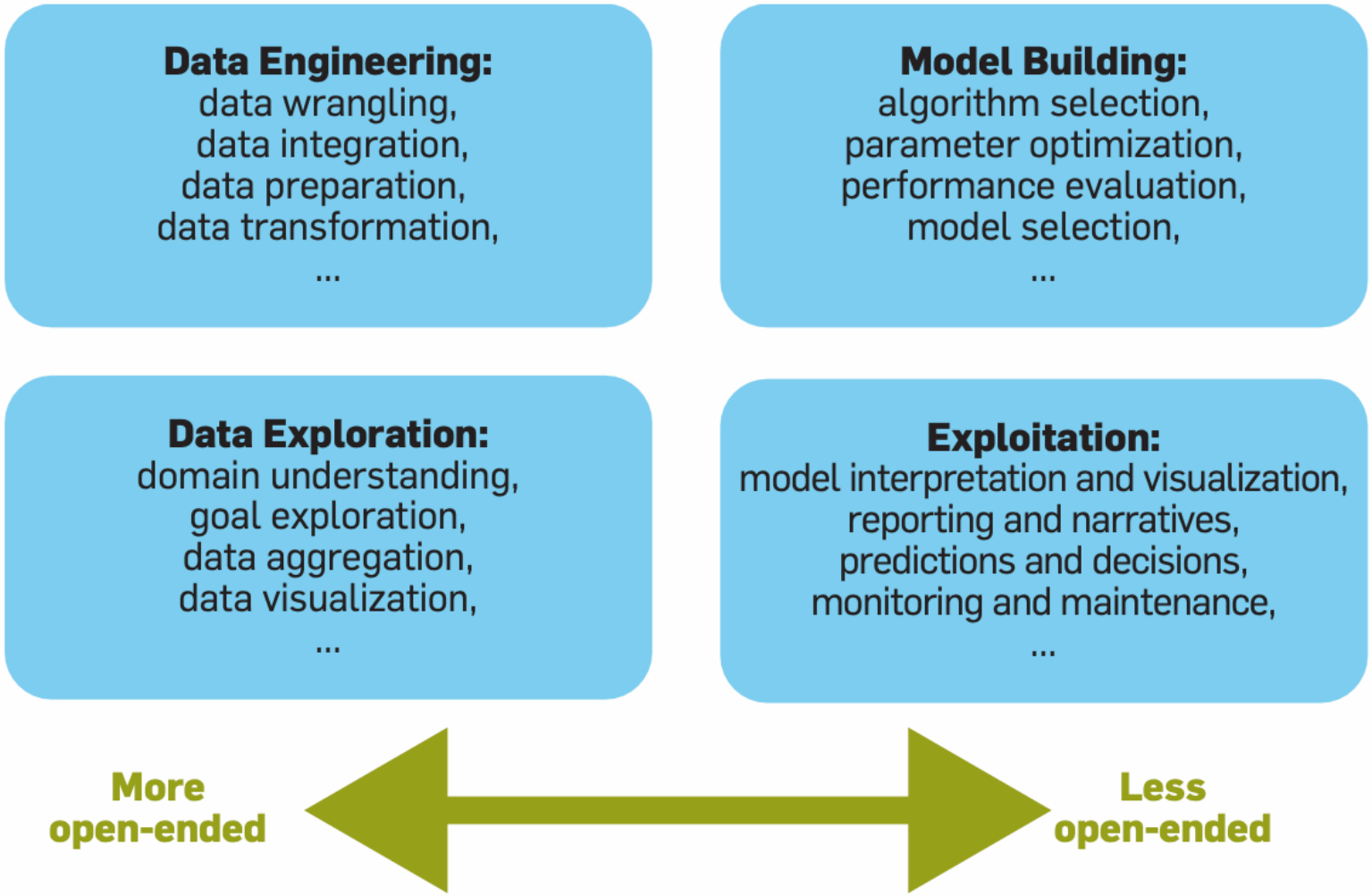
Data Exploration:
domain understanding,
goal exploration,
data aggregation,
data visualization,
...

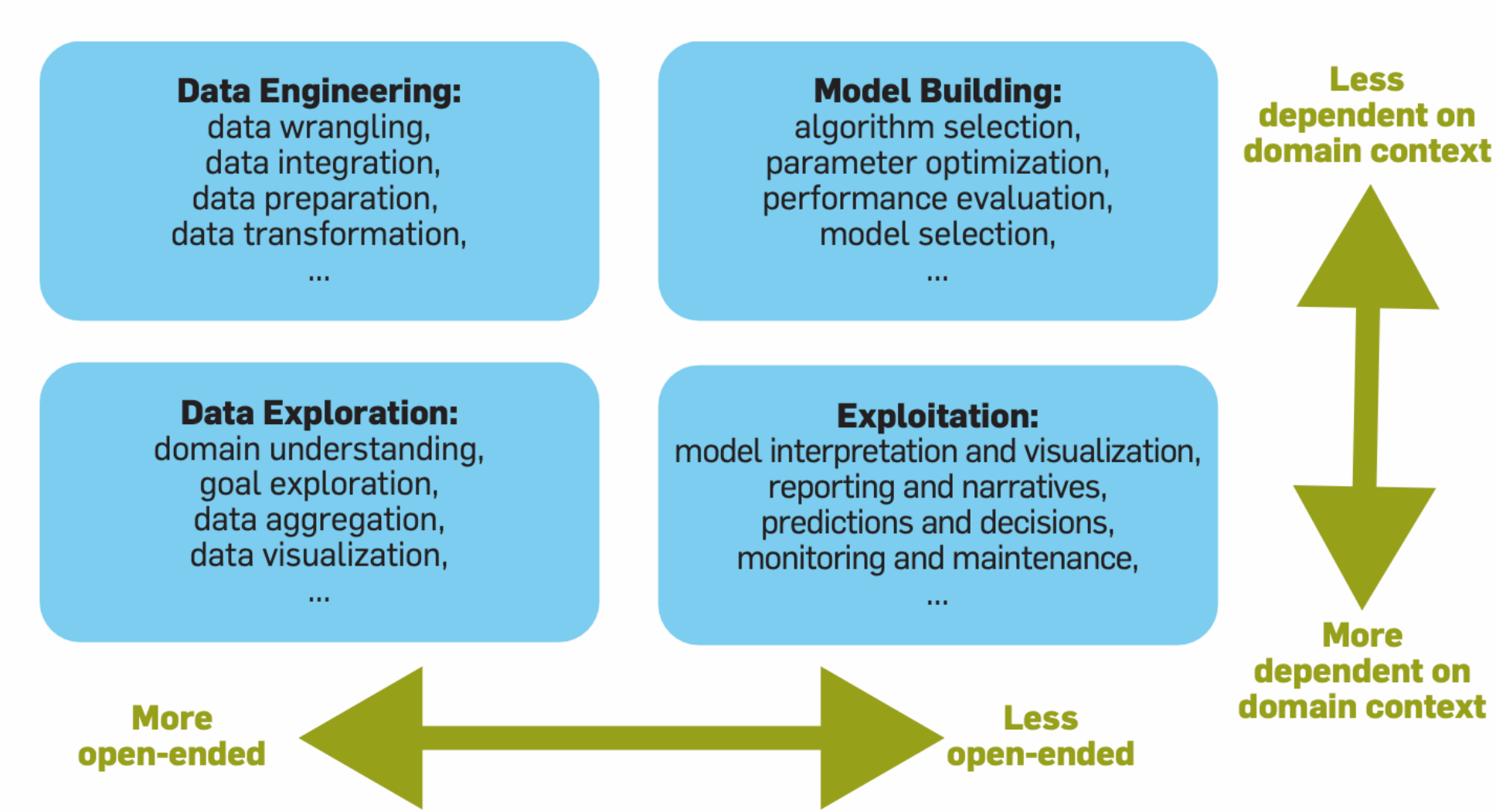
Exploitation:
model interpretation and visualization,
reporting and narratives,
predictions and decisions,
monitoring and maintenance,
...

TPS Exercise (T = done as part of homework)

Question:

Which of the four quadrants of data science are the most challenging to automate?


Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...
Model Building:
algorithm selection,
parameter optimization,
performance evaluation,
model selection,
...
Data Exploration:
domain understanding,
goal exploration,
data aggregation,
data visualization,
...
Exploitation:
model interpretation and visualization,
reporting and narratives,
predictions and decisions,
monitoring and maintenance,
...
More open-ended
Less open-ended


Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...
Model Building:
algorithm selection,
parameter optimization,
performance evaluation,
model selection,
...
Data Exploration:
domain understanding,
goal exploration,
data aggregation,
data visualization,
...
Exploitation:
model interpretation and visualization,
reporting and narratives,
predictions and decisions,
monitoring and maintenance,
...
Less dependent on domain context
More dependent on domain context
More open-ended
Less open-ended

TPS Exercise (T = done as part of homework)

Question:

What are the main challenges in data engineering?

TPS Exercise (T = done as part of homework)

Question:

What are the main challenges in data engineering?

Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...

Challenges:

- incomplete data, outliers
- inconsistent data formats
- data integration
- un- / semi-structured data

TPS Exercise (T = done as part of homework)

Question:

What are the three forms of automation relevant in the context of data science, and why do they all matter?

TPS Exercise (T = done as part of homework)

Question:

What are the three forms of automation relevant in the context of data science, and why do they all matter?

- mechanisation: fully automatic execution of a task
- composition: support for strategic sequencing of tasks, code- / workflow reuse
- assistance: identify (in)appropriate choices, make recommendations, …

Benefits of automation

Benefits of automation in data science / ML

- efficiency (less work)
- better results
- lower barrier to entry
- sustainability: less computational effort for same quality of results

TPS Exercise

Question:

What are pitfalls of automation?

TPS Exercise

Question:

What are pitfalls of automation?

- over-reliance, complacency
- errors that are subtle & difficult to detect
- cognitive bias towards observations, models, insights facilitated by given tools & systems

TPS Exercise

Question:

What are pitfalls of automation?

- over-reliance, complacency
- errors that are subtle & difficult to detect
- cognitive bias towards observations, models, insights facilitated by given tools & systems

What can be done to address these problems?

TPS Exercise

Question:

What are pitfalls of automation?

- over-reliance, complacency
- errors that are subtle & difficult to detect
- cognitive bias towards observations, models, insights facilitated by given tools & systems

What can be done to address these problems?

- focus on human-AI interaction
- monitoring AI systems
- look at the data! be sceptical! don’t drink the kool-aid!

Key concepts covered today:

- combined algorithm and hyperparameter optimisation (CASH)
- automatically constructing / optimising ML pipelines
  (leveraging general-purpose algorithm configurators)
- quadrants of data science:
  data exploration, data engineering, model building, exploitation
- forms of automation:
  mechanisation, composition, assistance
- benefits and pitfalls of automation

Key questions addressed in this module:

- **How good could an ML model be?**
  - Are we using the best possible ML method / model?
  - Have we configured and trained it in the best possible way?
  - Can we further improve performance?

Key questions questions addressed in this module:

- **How good could an ML model be?**
  - How can we ensure we are using a good ML method / model?
  - How can we configure and train it for optimised performance?
  - How can we further improve performance?

High-level learning goals:

Be able to …

- use standard tools and methods for selecting models and optimising their hyperparameters;
- explain AutoML concepts, tools and underlying methods at a technical level (Hoeffding races, grid/random search, Bayesian optimisation, NAS, …)
- be able to explain benefits and challenges in automating data science / ML

Further reading:

Chris Thornton, Frank Hutter, Holger Hoos, and Kevin Leyton-Brown:
Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms.
Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-13),
pp. 847-855, 2013

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, Frank Hutter:
Efficient and Robust Automated Machine Learning.
Proc. 28th International Conference on Neural Information Processing Systems (NIPS 2015),
pp. 2755–2763, 2013

Frank Hutter, Lars Kotthoff, Joaquin Vanschoren (editors):
AutoML: Methods, Systems, Challenges.
Springer, 2019

Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, Christopher K. I. Williams:
Automating Data Science. Communications of the ACM.
Vol. 65 No. 3, Pages 76-87, March 2022