Will Data Science abolish itself through automation?

AI Monday, 24th Feb 2020

paraboost DATA SCIENCE

AI FROM MARKGRÖNINGEN.

10+
Million Customers

2800+
Models

30+
Billion Predictions

paraboost
DATA SCIENCE

OUR MISSION.

DEMOCRATIZE ARTIFICIAL INTELLIGENCE

ACCELERATE MACHINE LEARNING PRODUCTIVITY

DOING THE NETFLIX APPROACH IN AI

BADER MODE. WOHNEN. LEBEN.

POKERSTARS

KLINGEL

S&P

Hotelwäsche ErwinMüller

baslerbeauty®

CONLEYS

BRIGITTE HACHENBURG exquisit

VEGA Immer eine Idee mehr

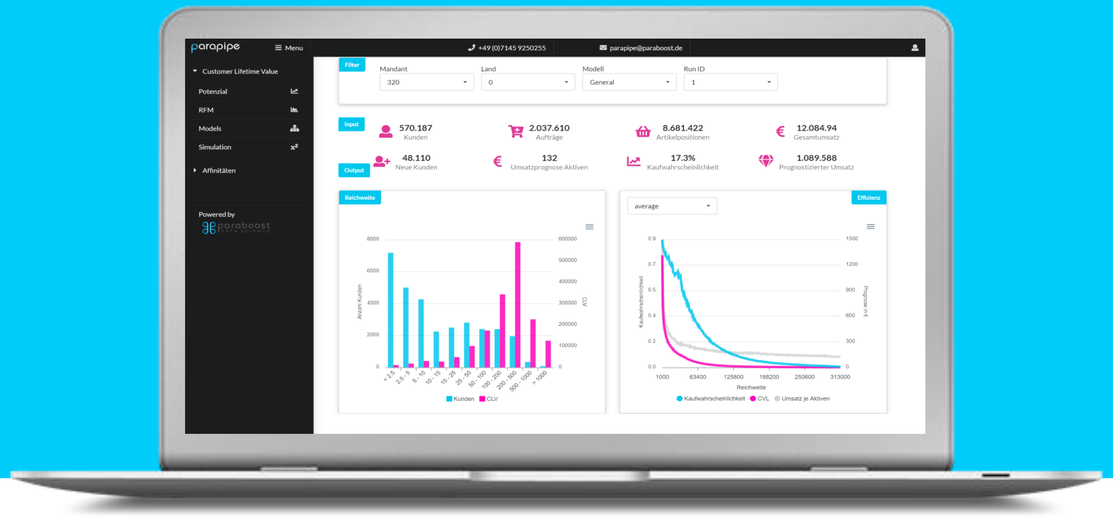
POKERSTARS CASINO

pulsiva

JOBELINE work · fashion · fun

OUR SOLUTION.

SELFLEARNING AI WITH OUSTANDING PERFORMANCE

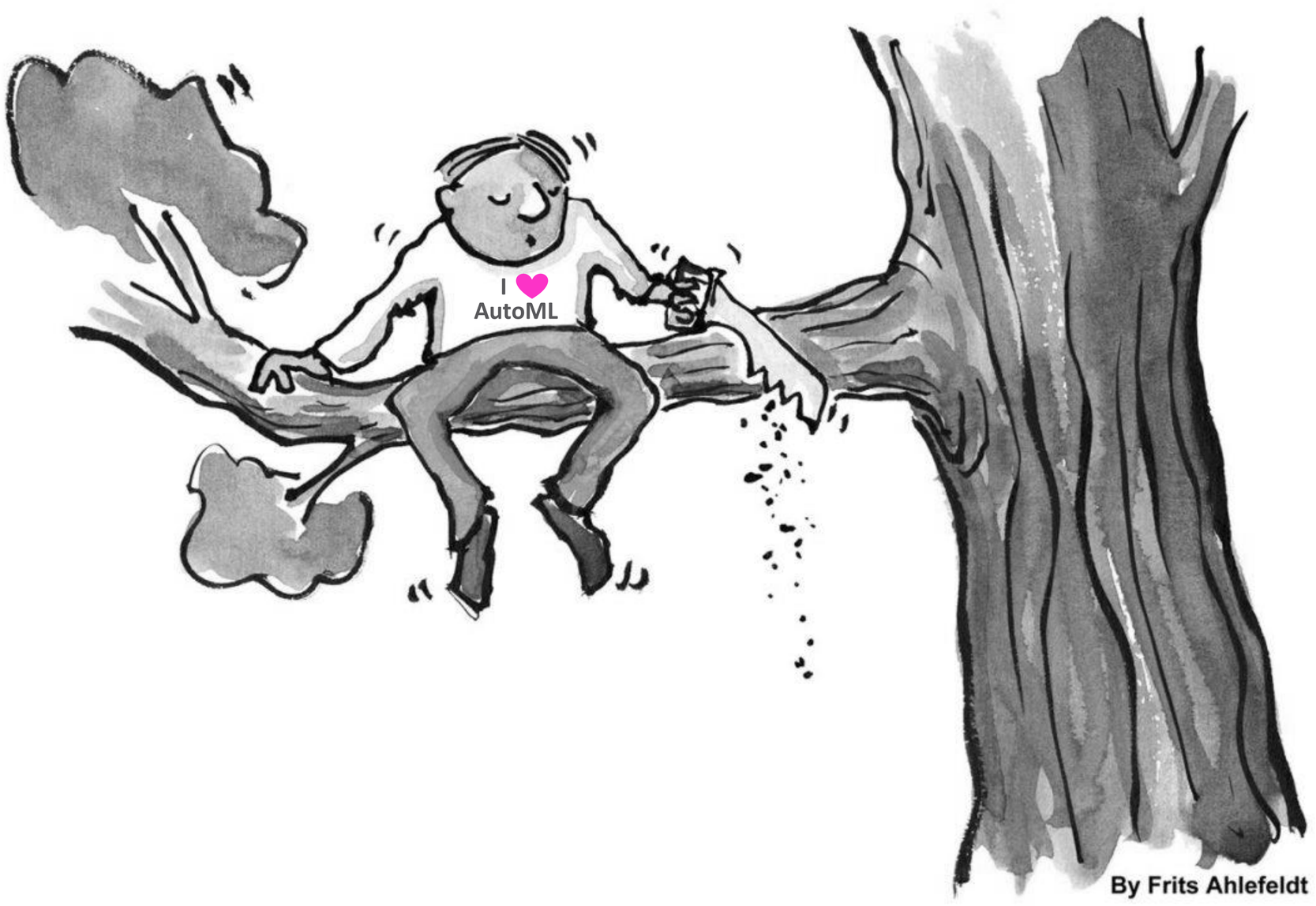


EXECUTABLE USE CASES

SAFE AI ENVIRONMENT

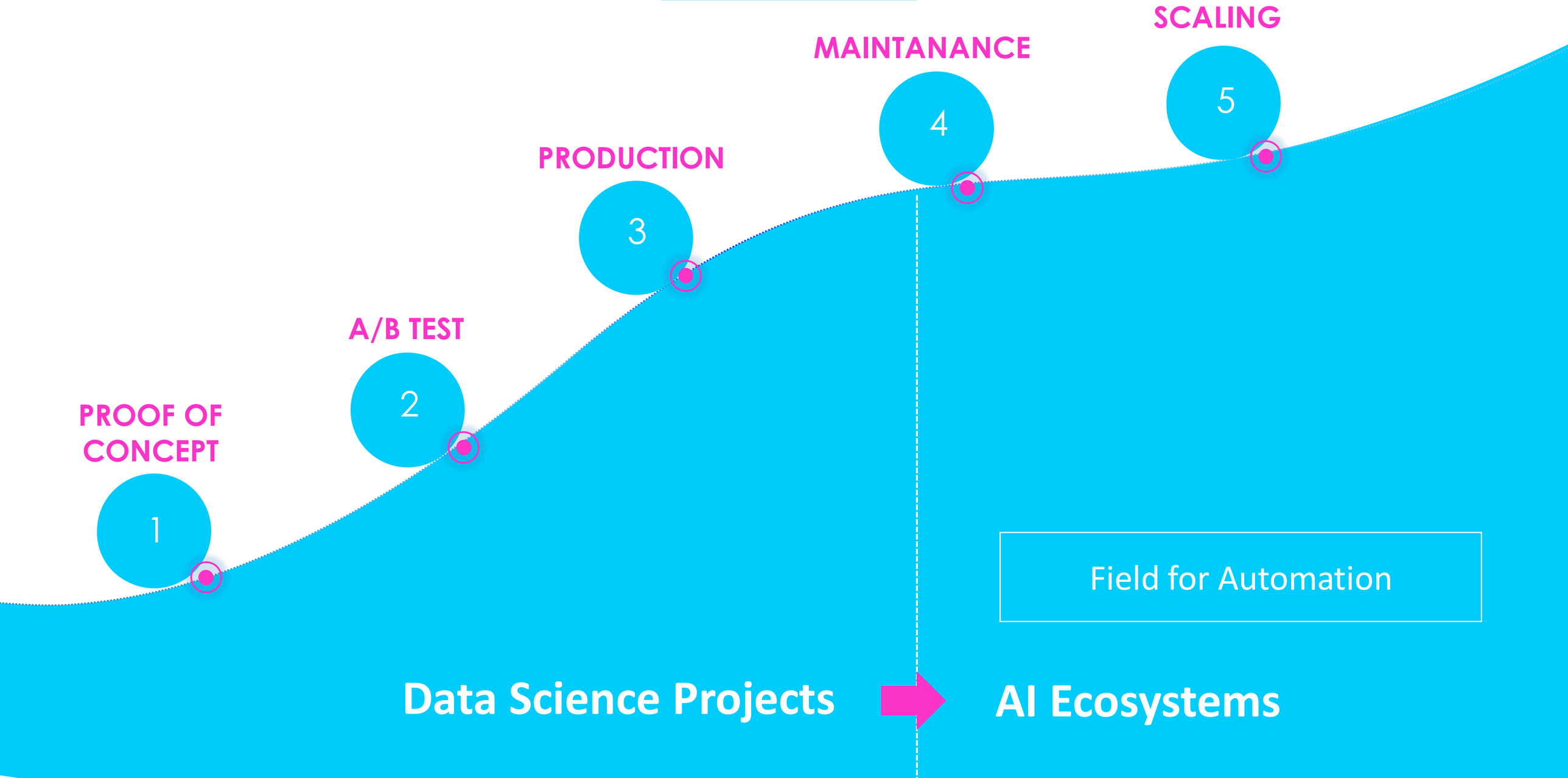
MASSIVE PARALLIZATION IN CPU & GPU

paraboost DATA SCIENCE



paraboost DATA SCIENCE

Maturity of Data Science Projects.

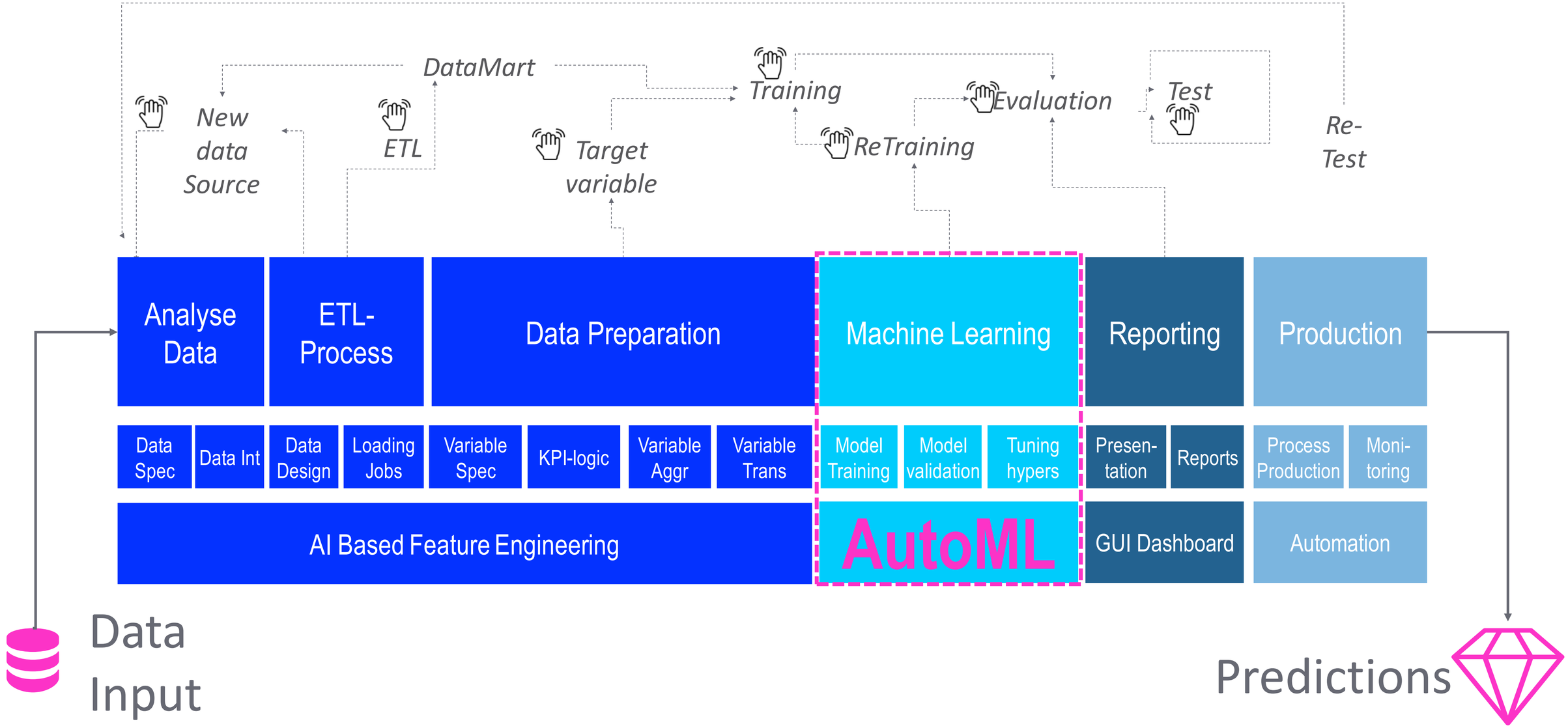


“

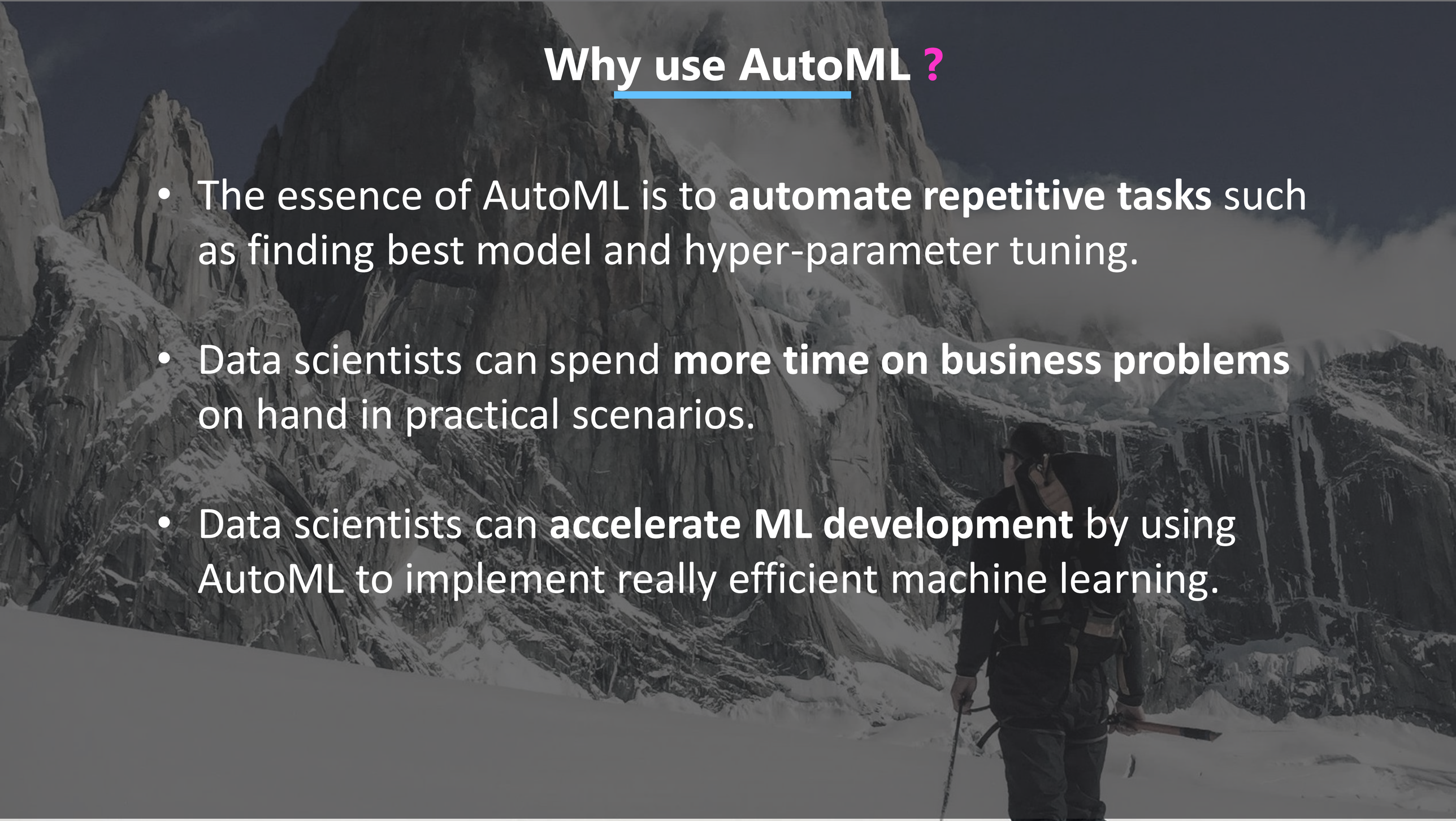
AI is like Sports-
If you want to get better, you need to **train** every day!

paraboost DATA SCIENCE

Data Science Process.



paraboost DATA SCIENCE



Why use AutoML ?

- The essence of AutoML is to **automate repetitive tasks** such as finding best model and hyper-parameter tuning.
- Data scientists can spend **more time on business problems** on hand in practical scenarios.
- Data scientists can **accelerate ML development** by using AutoML to implement really efficient machine learning.

AutoML - Under the hood.

Sampling techniques

How do I split training and validation dataset?

Outliers & Variable Selection

How do I remove outliers and select best attributes for model?

Algorithm Choice

What is the best algorithm for my model?

Hyperparameters & Ensembling

What are the best hyperparameters for my model and how can I stack them?

Leaderboard

What is the final best model that I choose?

Massive Experiments

paraboost DATA SCIENCE

```python
import h2o
from h2o.automl import H2OAutoML, get_leaderboard

h2o.init()

# Import a sample binary outcome train/test set into H2O
train = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_train_10k.csv")
test = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_test_5k.csv")

# Identify predictors and response
x = train.columns
y = "response"
x.remove(y)

# For binary classification, response should be a factor
train[y] = train[y].asfactor()
test[y] = test[y].asfactor()

# Run AutoML for 20 base models (limited to 1 hour max runtime by default)
aml = H2OAutoML(max_models=20, seed=1)
aml.train(x=x, y=y, training_frame=train)

# AutoML Leaderboard
lb = aml.leaderboard

# Optionally edd extra model information to the leaderboard
lb = get_leaderboard(aml, extra_columns='ALL')

# Print all rows (instead of default 10 rows)
lb.head(rows=lb.nrows)
```

```
# model_id                                                  auc    logloss     aucpr    mean_per_class_e
# --------------------------------------------------------  -----  ---------  -------  ------------------
# StackedEnsemble_AllModels_AutoML_20191213_174603     0.789844   0.551067  0.804672               0.31
# StackedEnsemble_BestOfFamily_AutoML_20191213_174603  0.789768   0.550906  0.805696               0.31
# XGBoost_grid__1_AutoML_20191213_174603_model_4       0.784698   0.55681   0.80312                0.32
# XGBoost_3_AutoML_20191213_174603                     0.784232   0.557749  0.802341               0.31
# XGBoost_2_AutoML_20191213_174603                     0.783533   0.555997  0.803189               0.32
# XGBoost_grid__1_AutoML_20191213_174603_model_3       0.782582   0.560218  0.800749               0.34
# GBM_5_AutoML_20191213_174603                         0.78219    0.558353  0.800234               0.31
# XGBoost_1_AutoML_20191213_174603                     0.781901   0.557944  0.801237               0.32
# XGBoost_grid__1_AutoML_20191213_174603_model_1       0.781648   0.561112  0.799203               0.31
# GBM_2_AutoML_20191213_174603                         0.777673   0.562514  0.796181               0.33
```

How to get started?

H2O AutoML

Auto-Sklearn

AUTO KERAS

TPOT

MLBox

AutoWeka

An Automated Data Mining Software Based on Weka

How will Automation impact Data Science?

1 Business

- Will have access to much more, much better, more updated predictive models in less time!
- Has more capacity to work on much more adhoc data science projects and use-cases.
- Needs to learn to **TRUST** models that are not build by humans.

2 Data Scientists

- Instead of building models data scientists will be building eco-systems for models.
- Will still be **very important** in the future, but not for repetitive tasks!
- **Collaboration** and **team development** becomes more important than ever before!

3 Technology

- AutoML can run thousands of experiments every day and require a lot of computational power.
- AutoML will need more GPU than CPU.
- High-End Gaming PC's with local GPU **can** be an alternative to cloud GPU.

Visit us!

Paraboost Marktplatz Office
Ostergasse 3
71706 Markgröningen

Contact us!

Paraboost GmbH

Contact person: Philipp Göller

Mobile: 0151 57868170
Email: p.goeller@paraboost.de
Web: www.paraboost.de

paraboost
DATA SCIENCE