HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

# Machine Learning for Accelerator Physics and Engineering

HELMHOLTZ AI

Peter Steinbach, Helene Hoffmann, Steve Schmerler, Sebastian Starke

*HZDR / ARD Workshop, Sep 24, 2021*

# Your Help Needed 🙏

## Science is about experiments ...

# Please open this online spreadsheet (PW: 1234)

In the spreadsheet, enter the following about yourself there:

T

- your gender (as 0 if female, 1 else)
- your shoesize
- your weight
- your height

**If you feel uncomfortable, feel free to ...**

make some of the numbers up or modify by $+1$ or $-1$ so that data cannot be tracked back to you!

# AI?

# A short (recent) history of AI ...

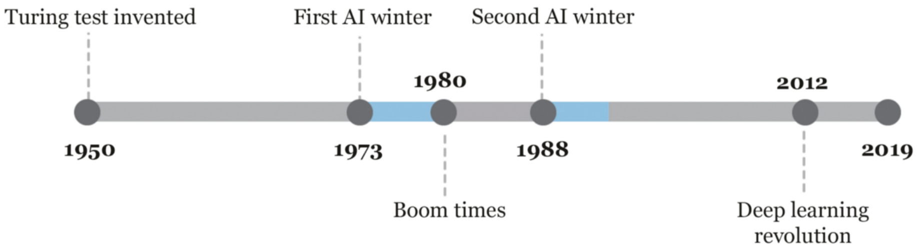
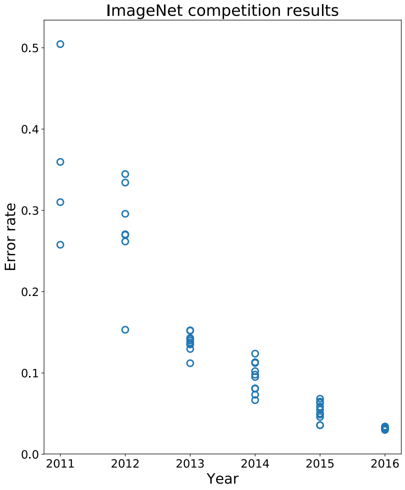



Figure: from Sebastian Schuchmann History of the first AI Winter

❄ failure (AI winters) and 🥂 success (AI Boom) alternate

🔭 mostly connected to high expectations

# The Imagenet Moment 2012 [5]




taken from Wikipedia:ImageNet

- curated database of ”images” and labels
- 15M images in 21k synonym classes
- 2017 (last): 2.25% classification error

# AlphaFold2 [4]

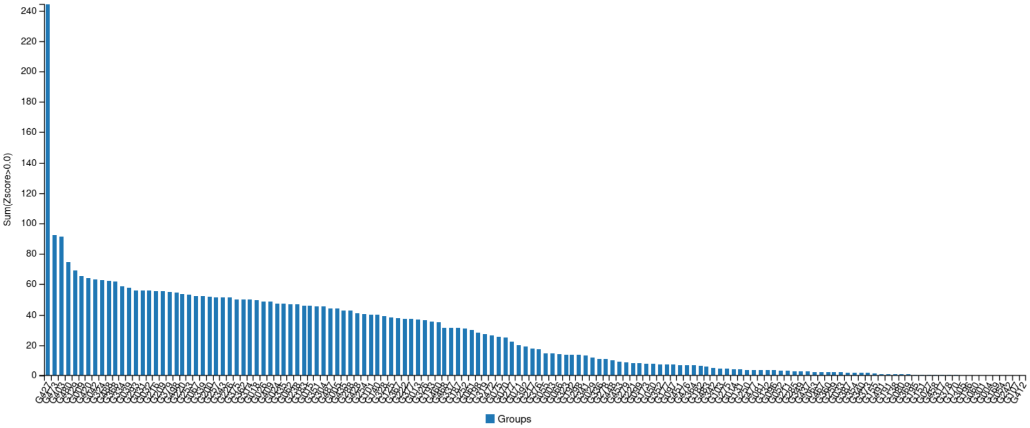



Figure: CASP14 results 2020

# Where are we today?

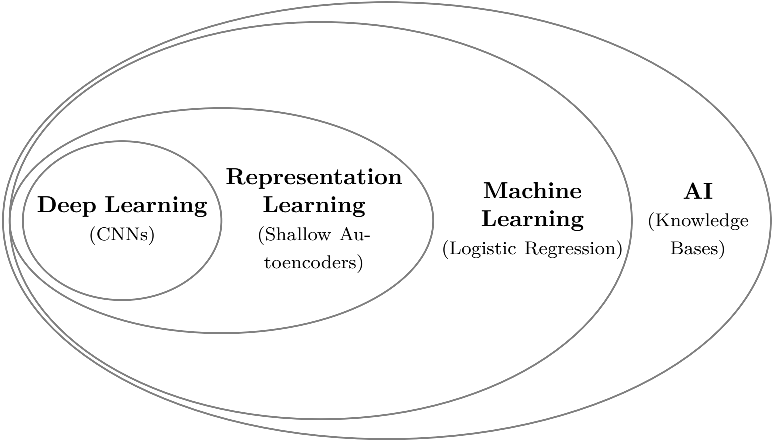



Figure: adopted from I. Goodfellow, Deep Learning, MIT Press [3]

# Helmholtz AI

# What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler

# What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026

# What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026
- 12 M€ per year (people + projects)

# What is Helmholtz AI?

- initiative by President of the Helmholtz Association, Prof. Otmar D. Wiestler
- running over 7 years, 2019 - 2026
- 12 M€ per year (people + projects)
- central installation in Munich (universities and Helmholtz center)

# Two Funding Lines

## Helmholtz AI Projects

unsplash.com:Glenn Carstens-Peters

- current call open until Dec 1, 2021
- max. 3 years, max. 200k € (must be matched)

## Helmholtz AI Vouchers

unsplash.com:Dominik Scythe

- voucher submissions open anytime
- get in touch first: consultant-helmholtz.ai@hzdr.de

# Helmholtz AI Local Unit For Matter At HZDR

Figure: Nico Hoffmann, YIG Lead

Figure: Peter Steinbach, Consultant Lead

# Helmholtz AI Consultant Team at HZDR




- reproducible automated (ML) pipelines
- inverse problems & generative modelling
- (image) denoising
- anomaly detection
- regression & pattern recognition (object localisation, image segmentation)
- aspects of trustworthy ML (uncertainties, robustness and interpretability)

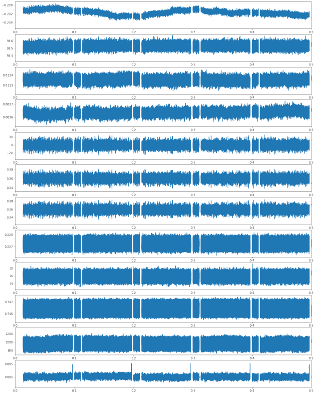
# Past and Present Vouchers

# Anomaly Detection at EUXFEL ...

## Can we automatically detect when damage happens to the synchronization laser?

Data:

- 10 s time snippets
- sensor output of healthy laser (training data)
- sensor output of damaged laser of same type (for testing)



# Our method: Feature extraction & Clustering

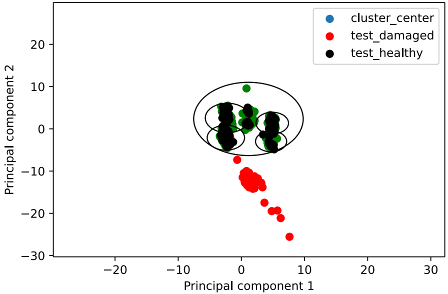
## Feature extraction:

- requirement: low dimensional representation of time-courses
- here: used `tsfresh`-package
  - simple features: mean, min, max, ...
  - more sophisticated features: fft-coefficients, entropy, absolute energy, ...

## Clustering:

- Principal Component Analysis
- kNN-Clustering

## Results:



# COSY simulation: MAD-X package @ FZJ [1]

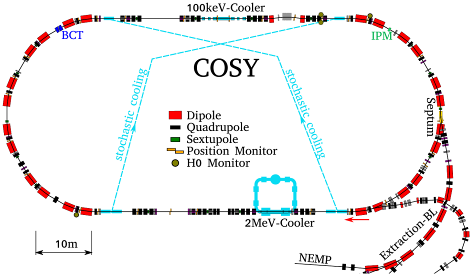
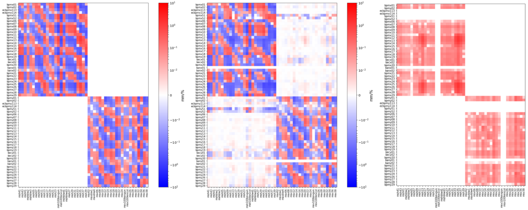


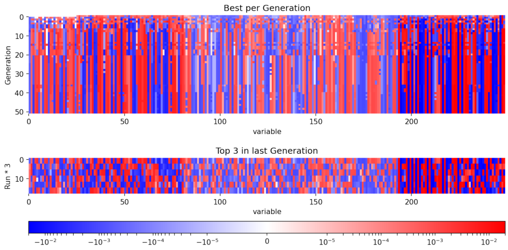

Figure 1: COSY lattice diagram. 184m circumference, split and extended 6-segment symmetry.



ORM simulation w/ initial params (left), exp. (middle),

- **goal:** improve simulation parameters
- simulation $f$ with 1500 parameters $\vec{x}$: $\hat{\mathbb{Y}} = f(\vec{x})$
- simulation output $\hat{\mathbb{Y}}$: 1 Orbit Response Matrix (ORM) with 3149 entries (only upper left+lower right used), 2 tune values, runtime $\sim$ 1 sec (fast!)
- optimization: fix most of $\vec{x}$ by measurements, **optimization goal: 249 free params** $\vec{x}' \in \mathbb{R}^{249}$
- data set $y$: 5 ORMs, $5 \times 2$ tunes (very small data set, need strong physics-based model $f$ = MAD-X)

# Evolutionary algorithm optimization




$x'^*$ = best $x'$ per generation during opt run (top), final top 3 $x'$ from 6 runs

- `deap` framework https://github.com/DEAP/deap, uses "population" of possible $x'$
- $C^* = \|y - f(x'^*)\|$ not as low as expected
- **many params $x_i'^*$ hit their allowed range limits**
- repeat runs: very similar $C^*$ but different $x'^*$ (similar to neural network optimization!)
- our suggestion: improve population initialization, do run monitoring (convergence behavior), check simulation (not accurate enough?), check values of fixed params not contained in $x'$, loss landscape analysis

# Prediction of spintune deviations at COSY synchrotron

- measurements at COSY of the spin tune over a period of several days showed unexpected deviations over time
- many monitoring variables are measured simultaneously with the spintune
- Goal: understand causes of the deviation
- Approach: build (interpretable) machine learning models to predict spintune deviation

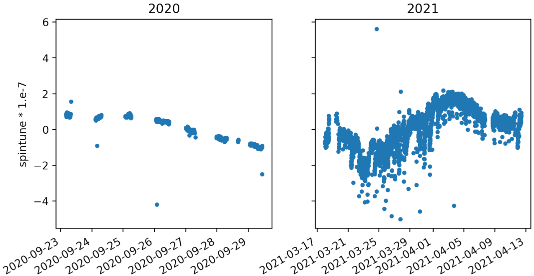



Figure: Deviation of spintune measurements

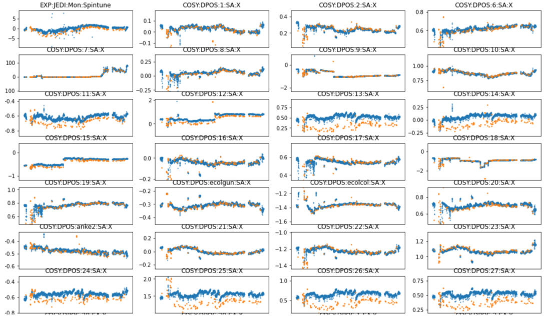



Figure: Features to predict spintune from

# Challenges and approaches

- **Challenges:**
  - data quality, outlier removal
  - partition of data into training, validation and test (should model interpolate or extrapolate)
- **Approaches:**
  - simple models (Linear regression, LASSO) do not give satisfactory performance, no obvious cause for spintune deviation has been identified yet
  - Kernel ridge regresion with laplacian kernel on PCA features looks promising, however interpretatability is limited

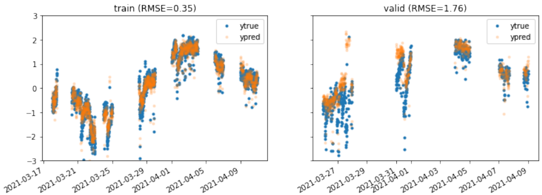



Figure: Linear regression prediction

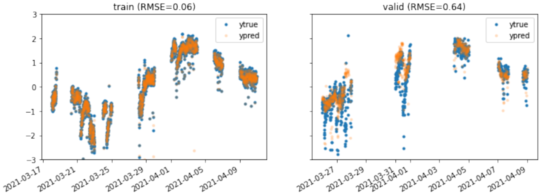
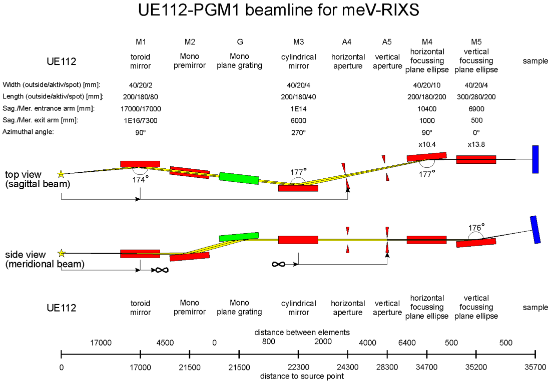
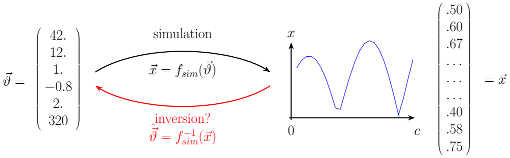



Figure: Kernel ridge regression with laplacian kernel on PCA features.

# Inverting a beamline simulation at BESSY ...



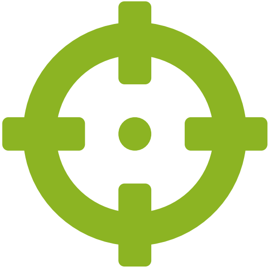




- goal: given a beamline profile (knife-edge scan), which beam control properties would result in this profile

# Lessons Learned

# Machine Learning needs a clear goal!

- narrow AI can solve/support many tasks
- needs mediation between domain experts and ML consultants
- started to use ML canvas [2]
- helped tremendously to structure projects



# Same method, different field!

- ML is software that can be tested! (Open Reproducible Science)
- talking about methods across disciplines
  - learn from others
  - get (professional) perspectives
- reference datasets can help to check feasibility (expectations)

# Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

## Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

- sorted chronologically and by camera

```
data
├── 2019
│   ├── 2019_04_01
│   ├── 2019_04_03
│   │   ├── AKO Bulli .................... camera name
│   │   ├── Lanex
│   │   └── SpEM
│   └── 2019_04_12
├── 2020
└── Shotlists
    ├── 2019
    └── 2020
```

# Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

- sorted chronologically and by camera
- CSV files and images in camera directory

```
data
└── 2019
    └── 2019_04_03
        └── AKO Bulli
            ├── set1_2019-04-03_11h-46m-54s_0.csv
            ├── set1_2019-04-03_11h-46m-54s_0_original.png
            ├── set1_2019-04-03_11h-52m-26s_1.csv
            └── set1_2019-04-03_11h-52m-26s_1_original.png
```

# Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

- sorted chronologically and by camera
- CSV files and images in camera directory
- important metadata encoded in the path

```
data
└── 2019
    └── 2019_04_03
        └── AKO Bulli
            ├── set1_2019-04-03_11h-46m-54s_0.csv
            ├── set1_2019-04-03_11h-46m-54s_0_original.png
            ├── set1_2019-04-03_11h-52m-26s_1.csv
            └── set1_2019-04-03_11h-52m-26s_1_original.png
```

# Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

- sorted chronologically and by camera
- CSV files and images in camera directory
- important metadata encoded in the path
- estimated 3.5 million files, 2.8 TB

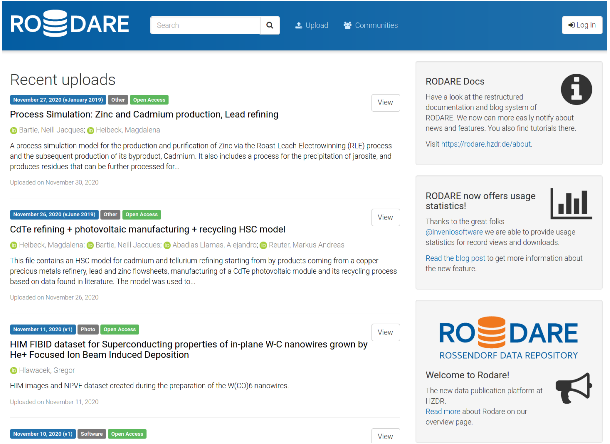
```
data
└── 2019
    └── 2019_04_03
        └── AKO Bulli
            ├── set1_2019-04-03_11h-46m-54s_0.csv
            ├── set1_2019-04-03_11h-46m-54s_0_original.png
            ├── set1_2019-04-03_11h-52m-26s_1.csv
            └── set1_2019-04-03_11h-52m-26s_1_original.png
```

# Data in the real world: Many Files in Many Directories (credits David Pape / HZDR)

- sorted chronologically and by camera
- CSV files and images in camera directory
- important metadata encoded in the path
- estimated 3.5 million files, 2.8 TB
- most files follow a template, but are often named and moved manually

```
data
└── 2019
    └── 2019_04_03
        └── AKO Bulli
            ├── set1_2019-04-03_11h-46m-54s_0.csv
            ├── set1_2019-04-03_11h-46m-54s_0_original.png
            ├── set1_2019-04-03_11h-52m-26s_1.csv
            └── set1_2019-04-03_11h-52m-26s_1_original.png
```

# FAIR principles [6]




HZDR Invenio open source software,
see also zenodo.org

Findable
Accessible
Interoperable
Reusable

BigData + FAIR = Necessity for ML

# Go FAIR with a csv (credits David Pape / HZDR)

### Findable

- use a public repository
- obtain unique global ID
- enrich metadata

### Accessible

- nobody to ask
- automated retrieval: data and metadata can be obtained by a freely implemented protocol

### Interoperable

- document based on standards (SI, datacite, ...)
- use established machine-readable formats (yaml, json, hdf5, tiff, ...)

### Reusable

- Choose a license!
- data meets community standards (description, i/o libraries, ...)

Automate and document the above as soon as possible!

# Demo Notebook: Predicting Shoe Sizes (doi:10.5281/zenodo.5541746)

1. share data publicly (doi:10.5281/zenodo.5541145)
2. download
3. open & check (pandas)
4. normalize, train and cross-validate (scikit-learn)
5. predict

# Summary

# Conclusion

- Helmholtz AI open for projects

# Conclusion

- Helmholtz AI open for projects
- already learned a lot about ML consulting projects

# Conclusion

- Helmholtz AI open for projects
- already learned a lot about ML consulting projects
- ML needs a concrete goal (expectations, testability, trustworthiness)

# Conclusion

- Helmholtz AI open for projects
- already learned a lot about ML consulting projects
- ML needs a concrete goal (expectations, testability, trustworthiness)
- ML works well on a FAIR dataset

# Conclusion

- Helmholtz AI open for projects
- already learned a lot about ML consulting projects
- ML needs a concrete goal (expectations, testability, trustworthiness)
- ML works well on a FAIR dataset

**Questions, Comments, Feedback or Concerns are highly welcome!**

# References

# References I


[1] I. Bekman and J. H. Hetzel. Cosy machine-model optimization. In *presented at the 12th Int. Particle Accelerator Conf. (IPAC'21)*, page 3375. JACoW Publishing, May 2021.

[2] Louis Dorard. Machine learning canvas, 2015.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

# References II

[6] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for

# References III

scientific data management and stewardship. 3(1):160018. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Publication characteristics;Research data Subject_term_id: publication-characteristics;research-data.

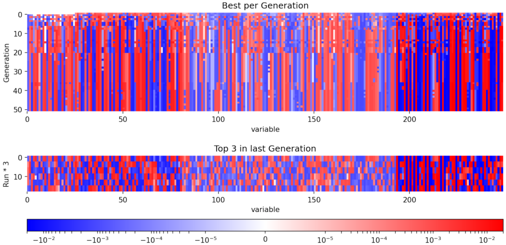
# Backup

# Evolutionary algorithm optimization

(blue/white is better)

$|meas - sim_{init}|$ $|meas - sim_{best}|$ $|meas - sim_{init}| - |meas - sim_{best}|$

ORM diff. to exp. from EA-optimized $x'^*$ (middle)




$x'^*$ = best $x'$ per generation during opt run (top), final top 3 $x'$ from 6 runs

- `deap` framework https://github.com/DEAP/deap, uses "population" of possible $x'$
- $C^* = \|y - f(x'^*)\|$ not as low as expected
- **many params $x_i'^*$ hit their allowed range limits**
- repeat runs: very similar $C^*$ but different $x'^*$ (similar to neural network optimization!)
- our suggestion: improve population initialization, do run monitoring (convergence behavior), check simulation (not accurate enough?), check values of fixed params not contained in $x'$, loss landscape analysis
- possible: loss landscape analysis (many optima, tune EA exploration behavior based on that)