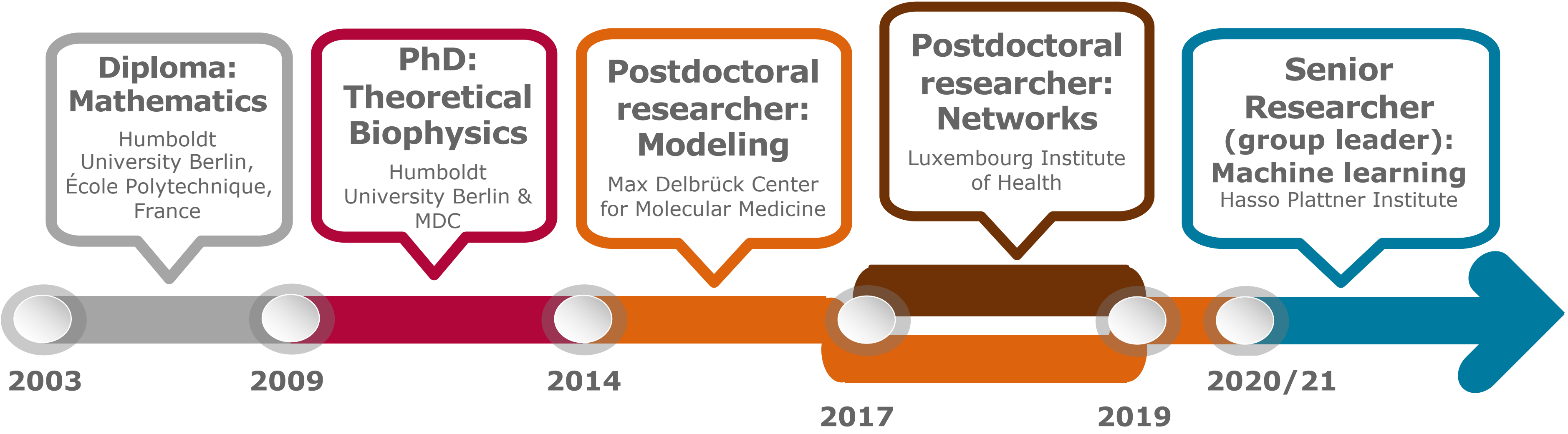
Integrating multi-layered data and prior knowledge into machine learning

Katharina Baum, Network-based data analysis
Hasso Plattner Institute for Digital Engineering, University of Potsdam

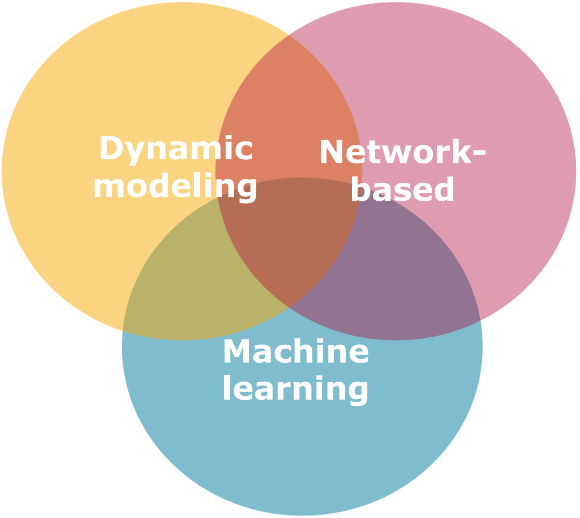
Spring School „Data Assimilation“, March 22, 2023

Katharina Baum

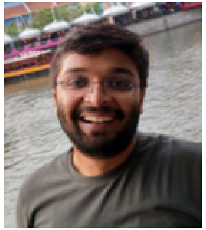


Overview: Network-based data analysis

- Draw from and combine different methods



Katharina Baum



Sukrit Gupta

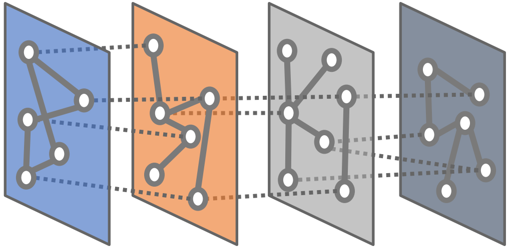
Pauline Hiort

Pascal Iversen

Theresa Hradilak

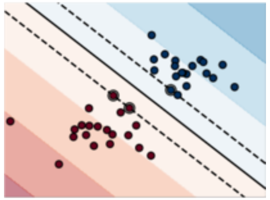
Master students
- Tim Garrels
- Pia Rissom
- Clemens Woest

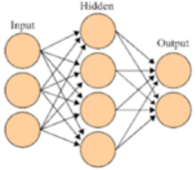
- Develop methods to analyze data across layers



ML, networks, dynamical models have different strengths and weaknesses

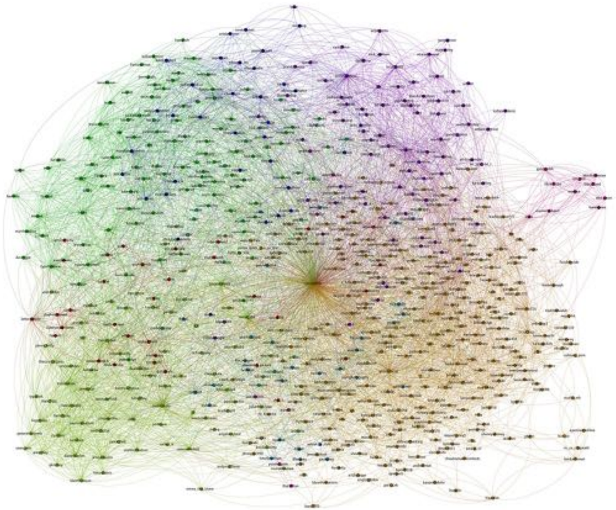
ML: predictions from unstructured data





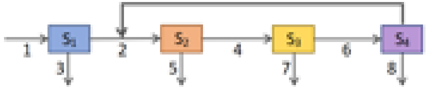
networks (graphs): capturing interactions

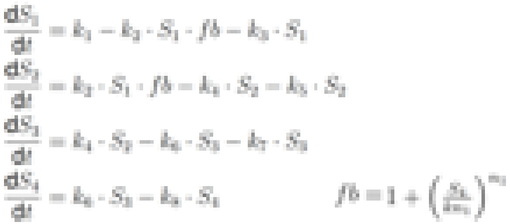
$$G = (V, E)$$

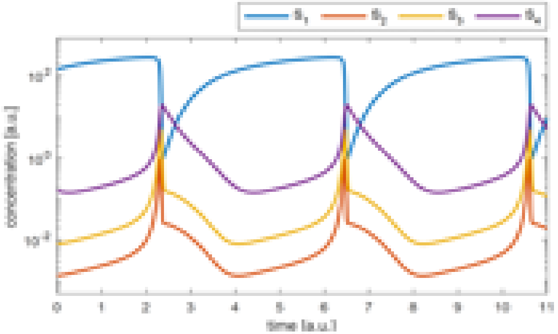


https://www.flickr.com/photos/speedoflife/6924482682; Andy Wang

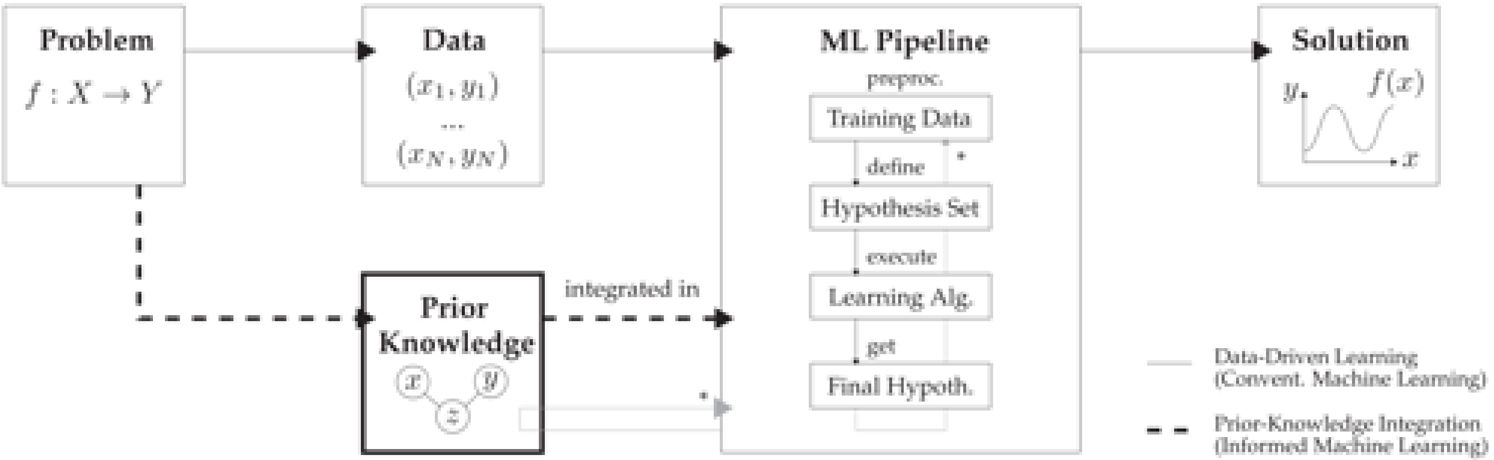
dynamical models: temporal properties



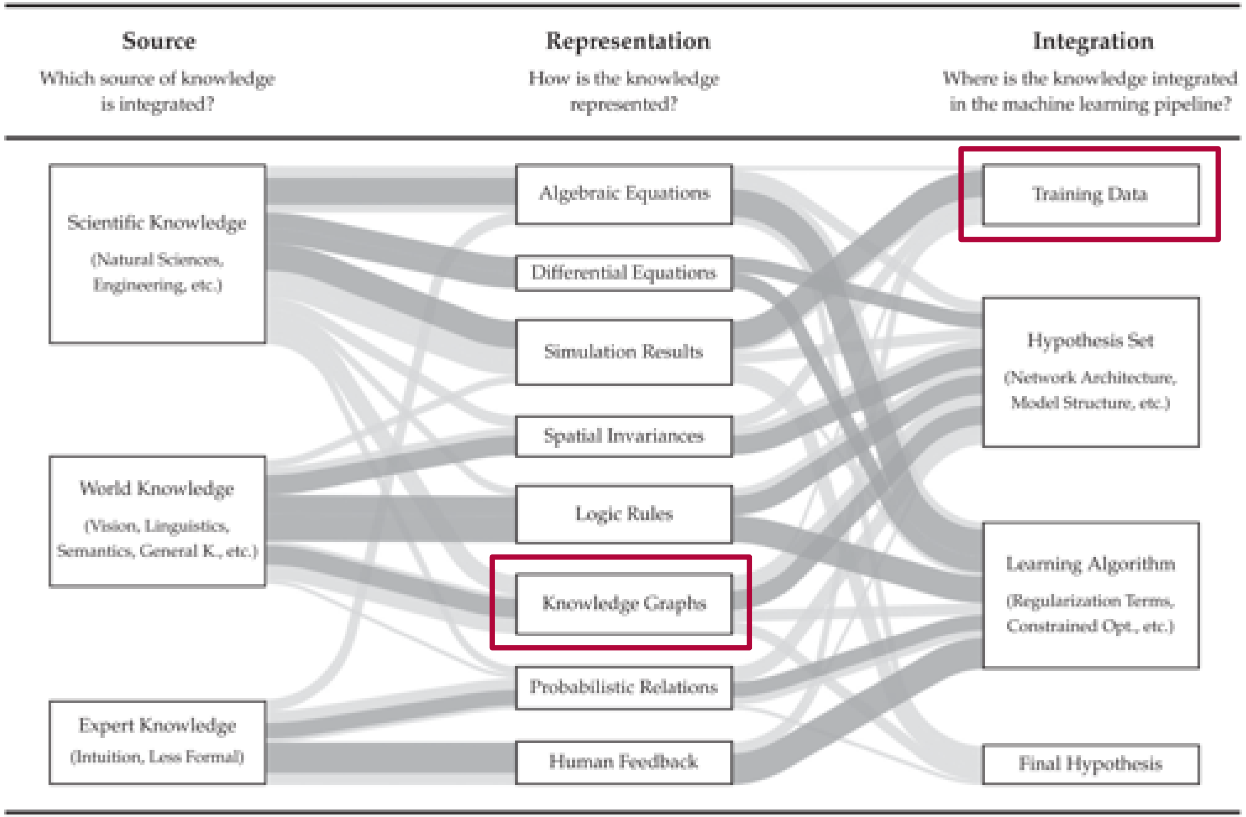




Informed machine learning von Rueden et al., 2023



Informed machine learning von Rueden et al., 2023, IEEE TKDE



Karniadakis et al., 2021 Nat Rev Phys

observation bias

inductive bias

learning bias

Including knowledge by feature engineering

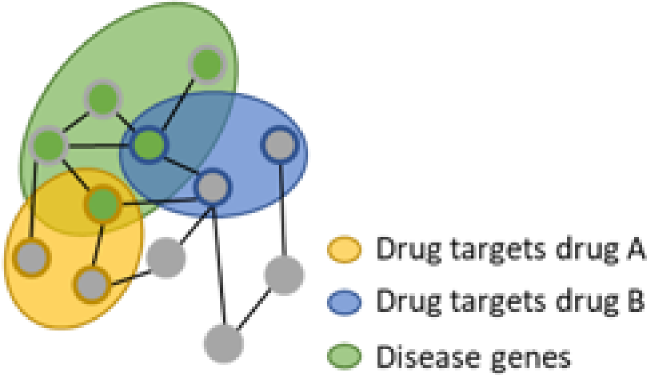
Prediction task:
given a disease D, and two drugs - classify the combination of drugs as good (approved, 1) or bad (adverse, 0)

$$f_D : X \times X \longrightarrow \{0,1\}$$

Pauline Hiort

molecular network

nodes: proteins, edges: their interactions



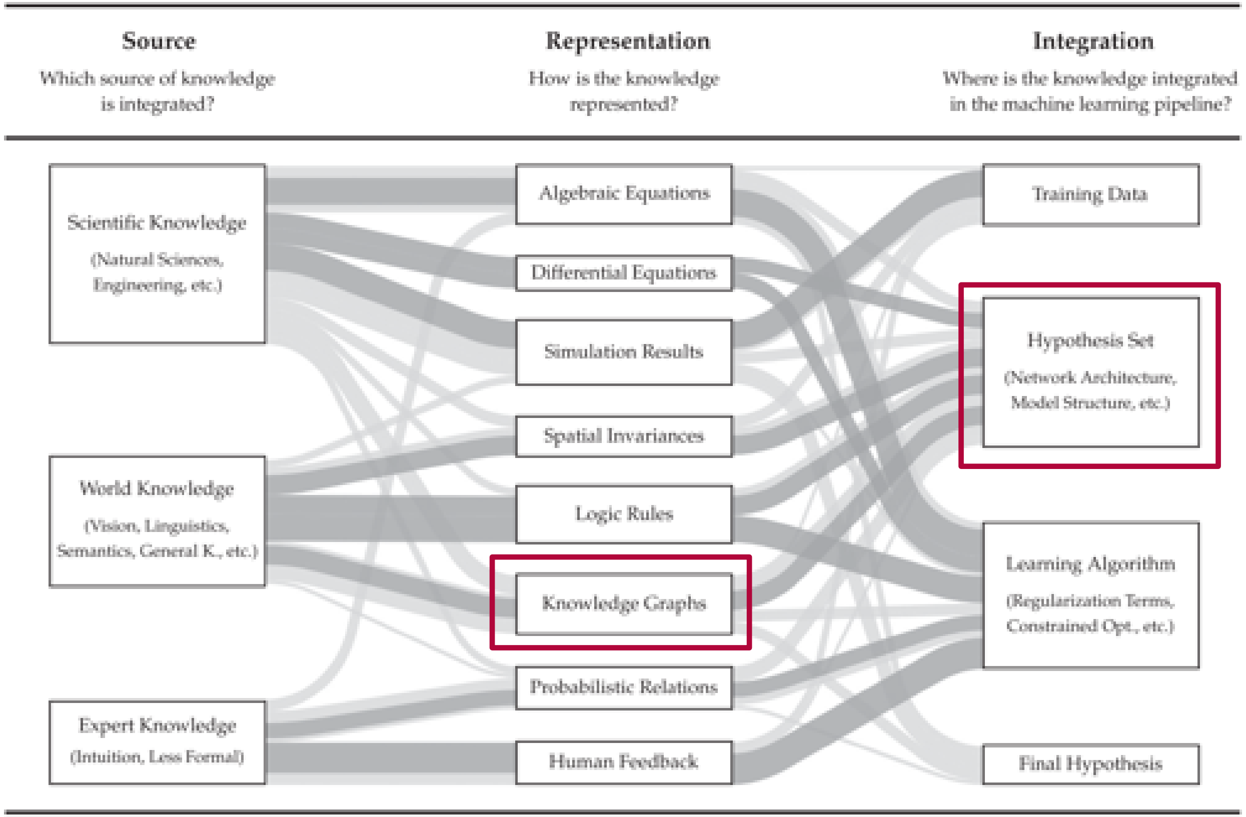
- prior knowledge/data
  (1) molecular network: protein-protein interactions
  (2) disease proteins
  (3) known targets of drugs
- infer features: distances

$$\widetilde{f_D} : \mathbb{R}^3 \longrightarrow \{0,1\}$$

- prediction with simple ML approaches (decision tree, SVM,...)

Informed machine learning

von Rueden et al., 2023 IEEE TKDE



Karniadakis et al., 2021, Nat Rev Phys

observation bias

inductive bias

learning bias

Graph-convolutional neural networks – bringing prior knowledge on proximity into ML

Kipf & Welling, ICLR 2017

We have a graph $G = (V, E)$ with $|V| = N$ with adjacency matrix A

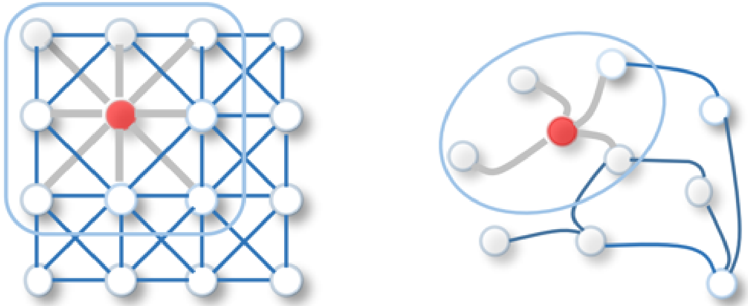
We have node feature vectors x_i of dimension F for $i = 1, \dots, N$, i.e. an NxF-dimensional node feature matrix X

The update rule for hidden layer l+1 is given by

$$H^{(l+1)} = f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

with $H^{(0)} = X$
$\hat{A} = A + I$
$\hat{D}$ degree diagonal matrix of $\hat{A}$
$W^{(l)}$ weight matrix of the lth neuronal layer
$\sigma(\cdot)$ nonlinear activation function

convolution over neighbours in the graph instead of over neighbouring pixels

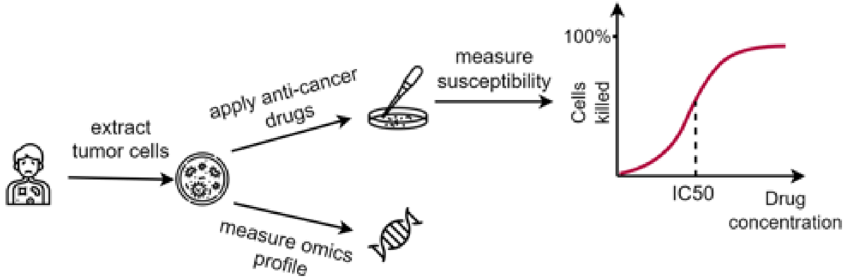


Wu et al, 2020, IEEE TNNLS

Including knowledge by using graph neural networks

Prediction task:
given a cell line, and a drug –
predict how strongly the cell line responds to the drug

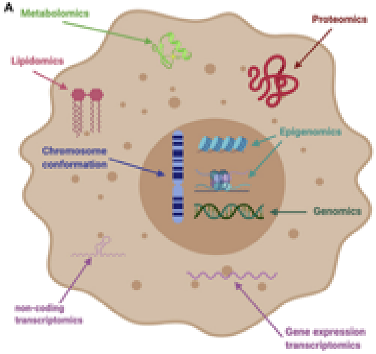
$$f: C \times X \longrightarrow \mathbb{R}$$



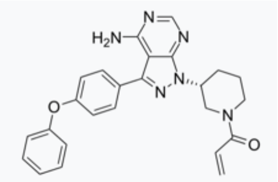
Pascal Iversen

- molecular properties of a cell line + number of features

Transcript-omics	Mutation	Methy-lation	Copy Nb Variation
17,737	30,333	14,726	20,669



- drug properties: targets, molecule structure, induced differential expression



Network-based data analysis

Katharina Baum

Including knowledge by using graph neural networks

- molecular properties of a cell line

Transcript-omics	Mutation	Methy-lation	Copy Nb Variation
17,737	30,333	14,726	20,669

- drug properties: targets, molecule structure, induced differential expression

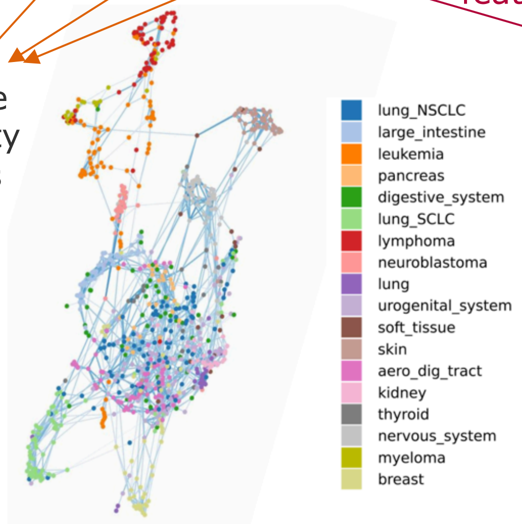
features

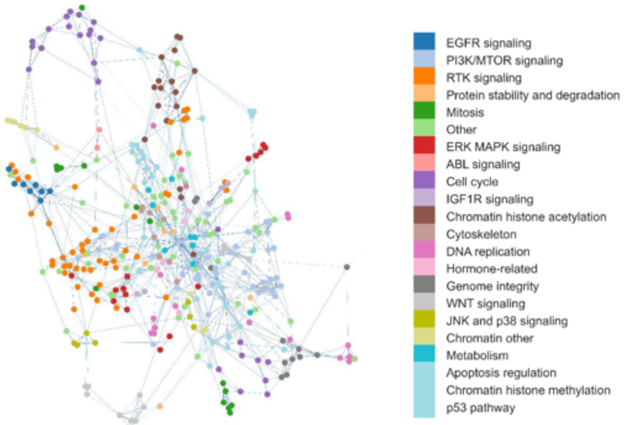
features

cell line similarity graphs

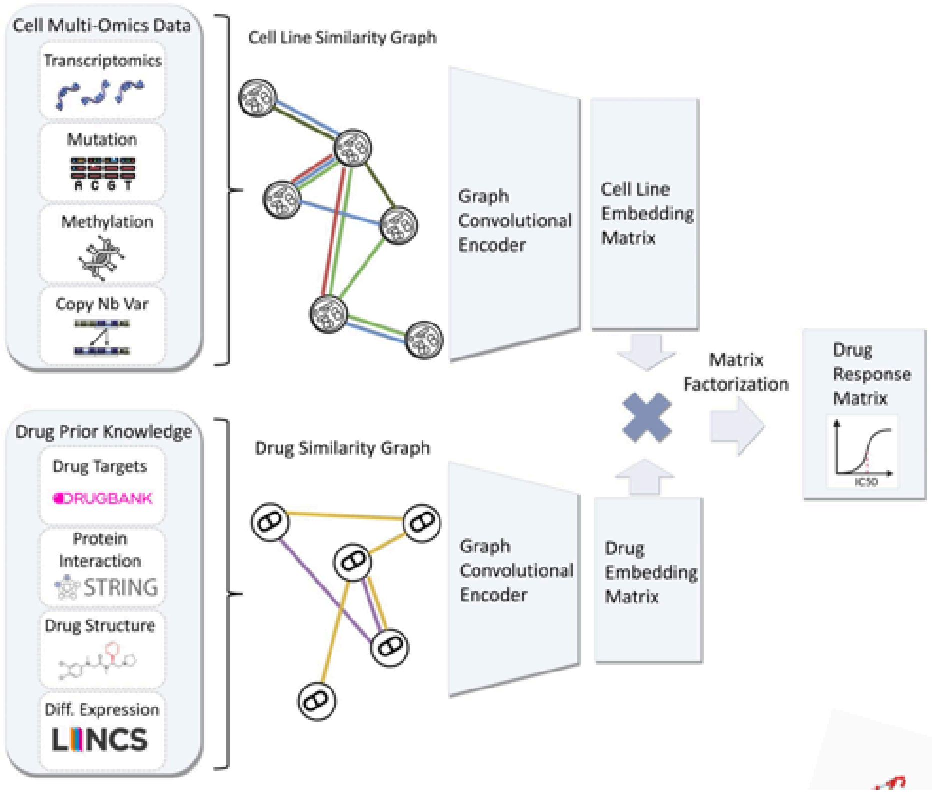
$$f: C \times X \longrightarrow \mathbb{R}$$

drug similarity graphs



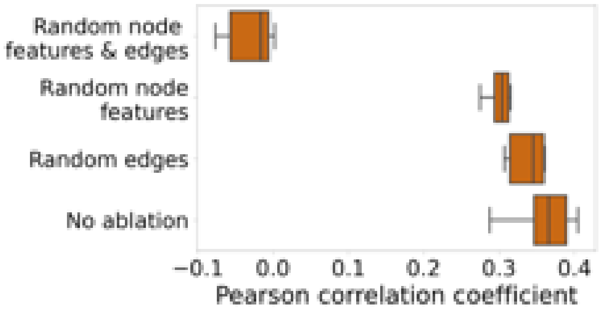


Drug response prediction with similarity graphs



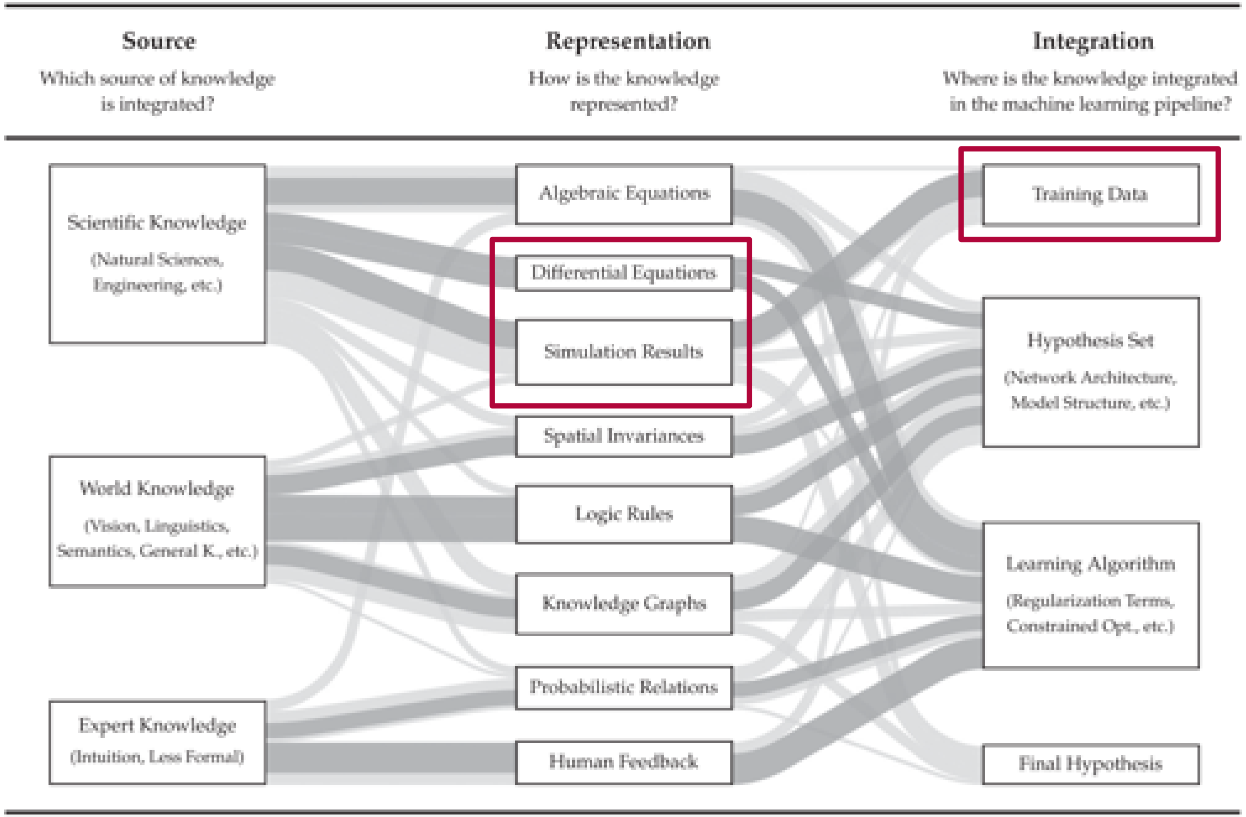
Model	Pearson	MSE
GCMF	0.36 ± 0.04	0.88 ± 0.03
U-GCMF	0.34 ± 0.06	0.90 ± 0.04
PaccMann[3]	0.35 ± 0.02	0.97 ± 0.10
Ridge	0.32 ± 0.02	0.94 ± 0.02
SRMF[4]	0.01 ± 0.02	1.11 ± 0.03

Ablation study



Informed machine learning

von Rueden et al., 2023 IEEE TKDE



Karniadakis et al., 2021, Nat Rev Phys

observation bias

inductive bias

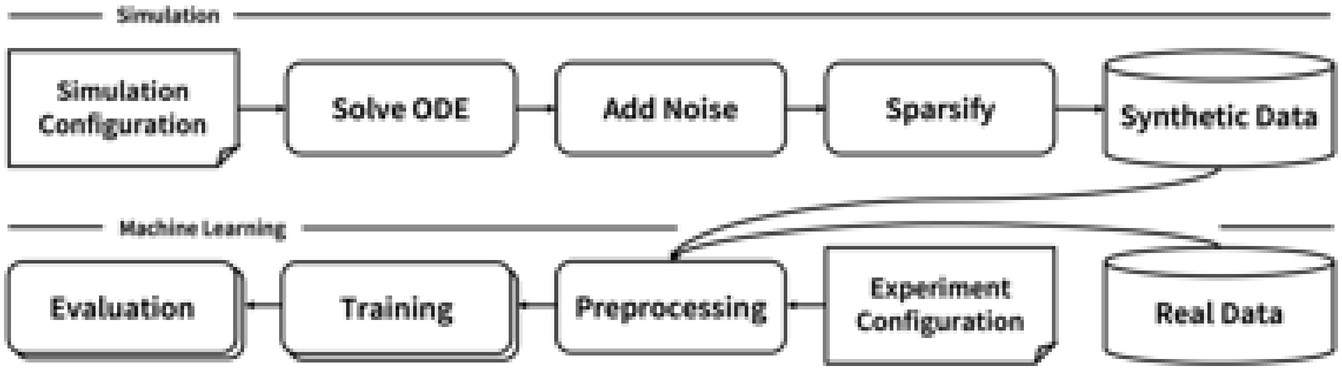
learning bias

Simulation-based Machine Learning

- Maximilian Kleissl, Benedict Heyder, Lukas Drews, Julian Zabbarov

co-supervision by Pascal Iversen

Idea: a toolbox for model-informed machine learning



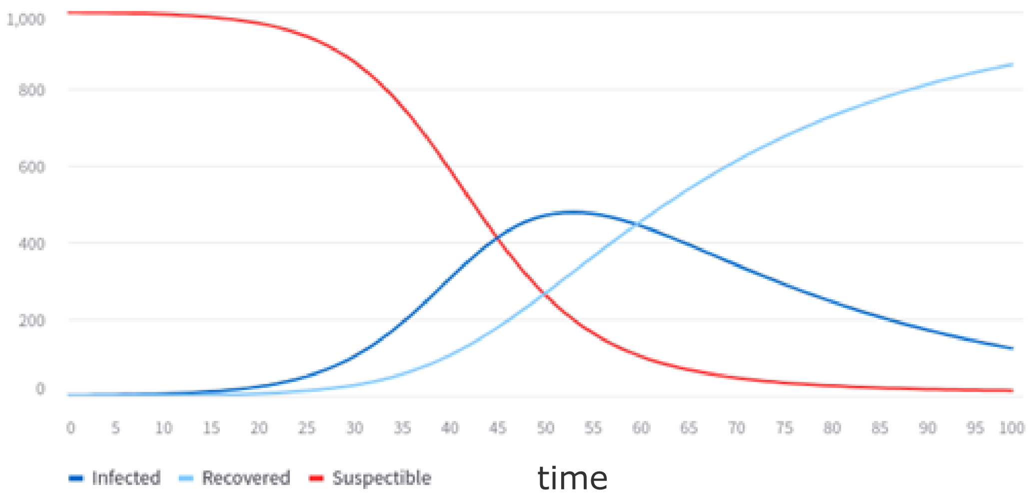
Simulation with SimbaML: SIR model

SIR: susceptible, infected, recovered

$$\frac{dS}{dt} = -\frac{\beta I S}{N},$$

$$\frac{dI}{dt} = \frac{\beta I S}{N} - \gamma I,$$

$$\frac{dR}{dt} = \gamma I,$$



config:

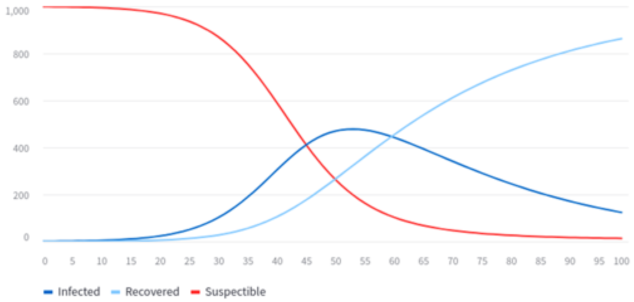
- ranges or distributions of initial conditions
- ranges or distributions of kinetic parameters
- solver, error, time series or steady state
- noise, constraints,…

Different noise options

$$\frac{dS}{dt} = -\frac{\beta IS}{N},$$

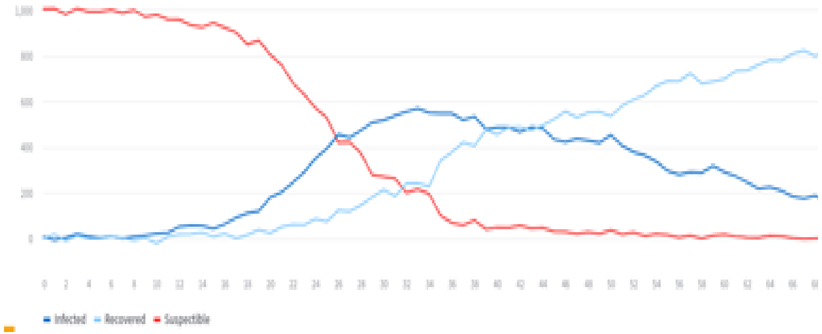
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I,$$

$$\frac{dR}{dt} = \gamma I,$$



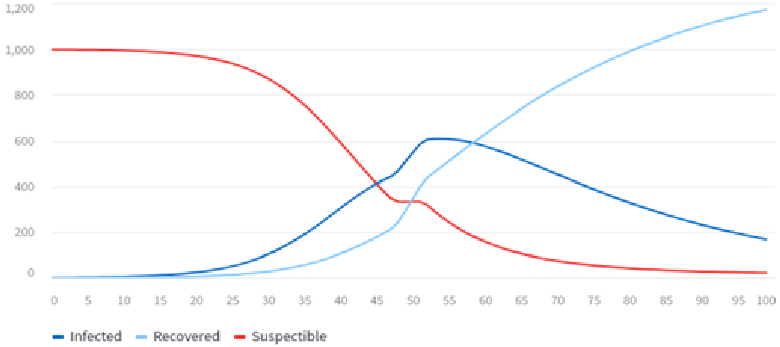
measurement errors

$$S(t) + \varepsilon_{S(t)}$$
$$R(t) + \varepsilon_{R(t)}$$
$$I(t) + \varepsilon_{I(t)}$$
$$\varepsilon_{\alpha} \sim \mathrm{N}(0, \sigma^2)$$

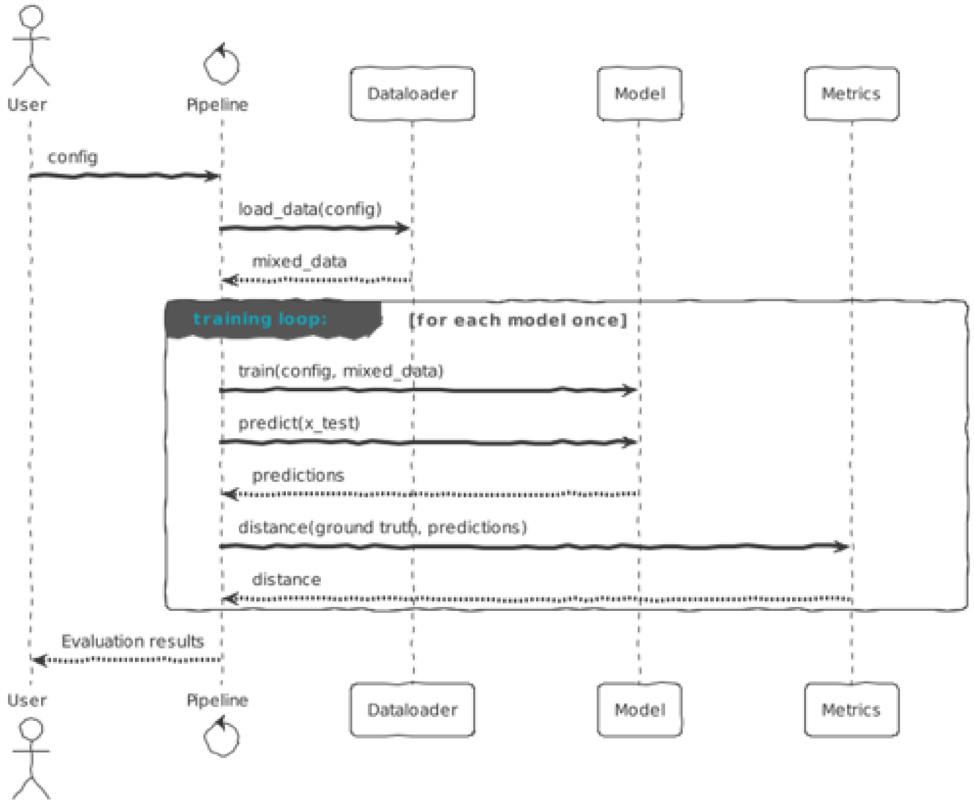


intrinsic errors

$$\frac{dS}{dt}(t) + \varepsilon_t$$



Machine learning with SimbaML



SimbaML supports ML models from
- Keras,
- PyTorch Lightning, and
- scikit-learn.

Potential application of SimbaML: model pre-training in sparse data situations

prediction task: time series forecasting of the distribution parameters for the number of new infections, x

$$f: \big(x(t-6), \dots, x(t)\big) \rightarrow (\mu(t+1), \dots, \mu(t+7), v(t+1), \dots, v(t+7))$$

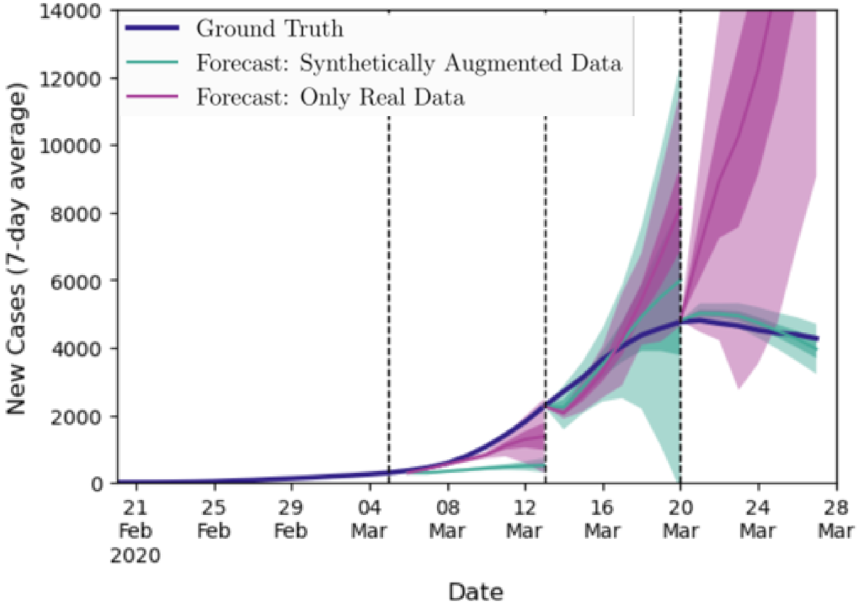
assuming

$\mathrm{x}(\mathrm{t}+\mathrm{k}) = \mu(t+k) + \varepsilon(t+k)$

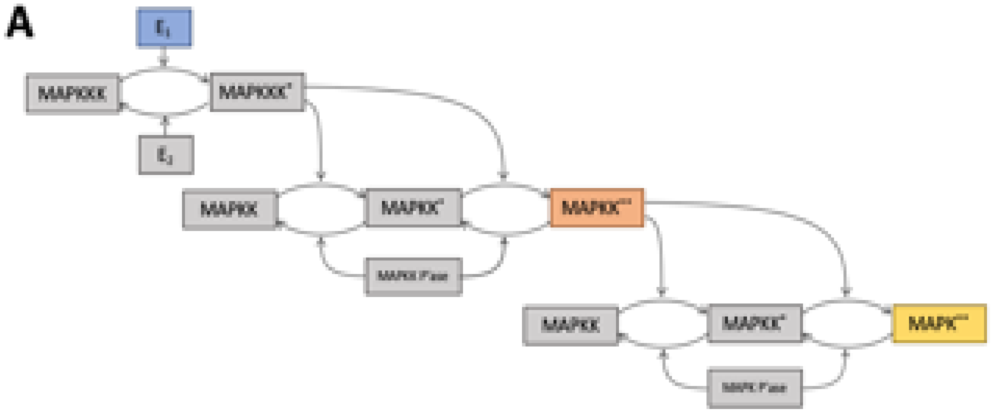
with $\varepsilon(t+k) \sim T(v(t+k))$

Can we supplement the model with simulated data from SimbaML?

- simulate time series using an SIR model



Potential application of SimbaML: Determine the best ML prediction model

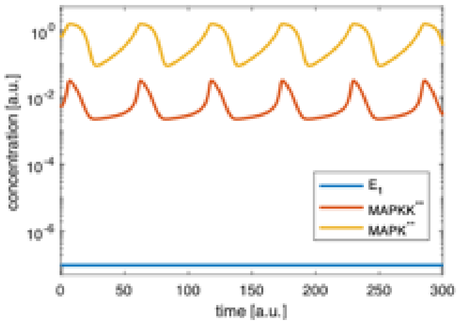


$$\frac{dS_8}{dt} = -k_{19} \cdot S_8 \cdot S_7 + k_{20} \cdot S_{17} + k_{24} \cdot S_{22}$$
$$\frac{dS_9}{dt} = k_{21} \cdot S_{17} - k_{22} \cdot S_9 \cdot S_{11} + k_{23} \cdot S_{22} - k_{25} \cdot S_9 \cdot S_7 + k_{26} \cdot S_{18} + k_{30} \cdot S_{21}$$
$$\frac{dS_{10}}{dt} = k_{27} \cdot S_{18} - k_{28} \cdot S_{10} \cdot S_{11} + k_{29} \cdot S_{21}$$
$$\frac{dS_{11}}{dt} = -k_{22} \cdot S_9 \cdot S_{11} + k_{23} \cdot S_{22} + k_{24} \cdot S_{22} - k_{28} \cdot S_{10} \cdot S_{11} + k_{29} \cdot S_{21} + k_{30} \cdot S_{21}$$
$$\frac{dS_{12}}{dt} = -k_{10} \cdot S_6 \cdot S_{12} + k_{11} \cdot S_{20} + k_{12} \cdot S_{20} - k_{16} \cdot S_7 \cdot S_{12} + k_{17} \cdot S_{19} + k_{18} \cdot S_{19}$$
$$\frac{dS_{13}}{dt} = k_1 \cdot S_1 \cdot S_3 - k_2 \cdot S_{13} - k_3 \cdot S_{13}$$
$$\frac{dS_{14}}{dt} = k_4 \cdot S_2 \cdot S_4 - k_5 \cdot S_{14} - k_6 \cdot S_{14}$$

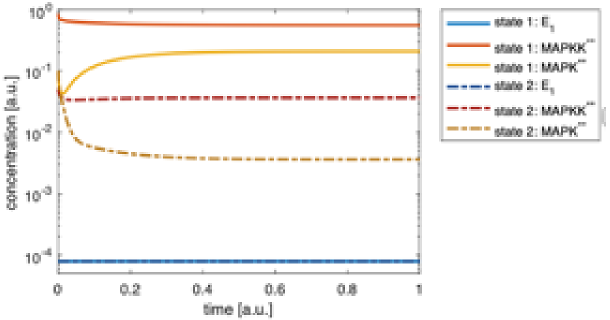
ordinary differential equation (ODE) model

(Huang & Ferrell 1996)

$$\frac{dS_1}{dt} = -k_1 \cdot S_1 \cdot S_3 + k_2 \cdot S_{13} + k_3 \cdot S_{13}$$



$- k_{25} \cdot S_9 \cdot S_7 + k_{26} \cdot S_{18} + k_{27} \cdot S_{18}$



Network-based data analysis

Katharina Baum

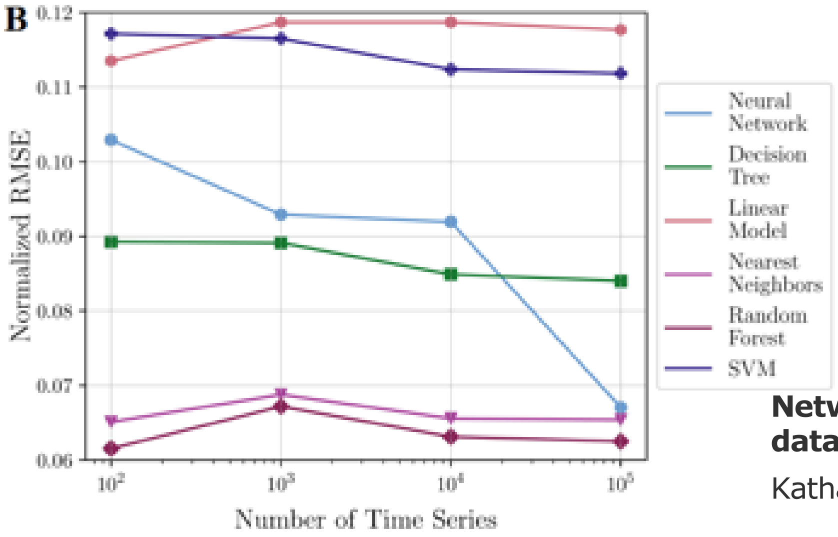
Potential application of SimbaML: Determine the best ML prediction model

prediction task: time series forecasting of a single observed variable of the system, x

$$f: \left(x(t-4), \dots, x(t)\right) \rightarrow \left(x(t+1), x(t+2), x(t+3)\right)$$

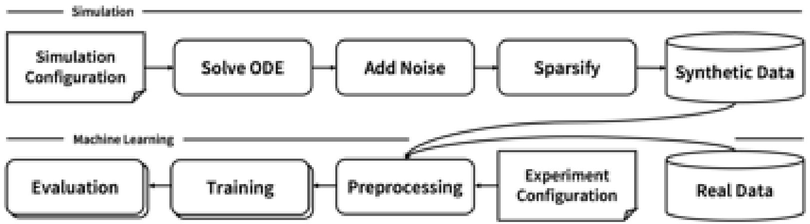
Which ML model performs best with the given amount of training data?

- synthesize different numbers of time series (20 time steps length) with SimbaML on the basis of the MAPK ODE model
- train using different ML models



Outlook

SimbaML



- examination of effects of different assumptions of noise
- explore transfer learning
- benchmark methods of explainability

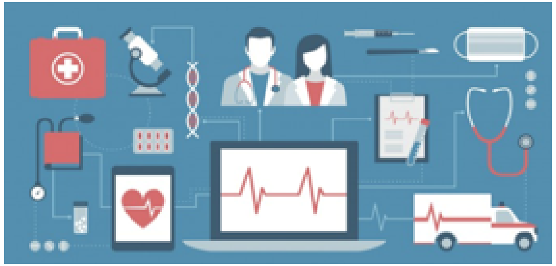
- assess other informed ML approaches
- role of uncertainty, active learning

In general:
transfer to clinically relevant settings
(1) personalized predictions
(2) transfer learning approaches
(3) explainable predictions
(4) include additional data

Genomics england

HPI·MS Hasso Plattner Institute for Digital Health at Mount Sinai

ICAHN School of Medicine at Mount Sinai



Thank you!

Katharina Baum, Network-based data analysis,
Hasso Plattner Institute for Digital Engineering, University of Potsdam

Spring School „Data Assimilation", March 22, 2023