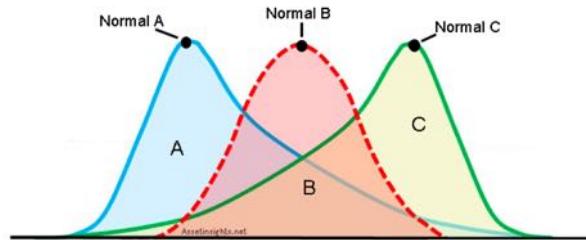


Statistical Relational Learning in Healthcare*

*Thanks for Sriraam Natarajan (UT Dallas)
and many others for all the great collaborations



Kristian Kersting



Getting deep
systems that reason
and know when they
don't know

Responsible AI
systems that explain
their decisions and
co-evolve with the
humans

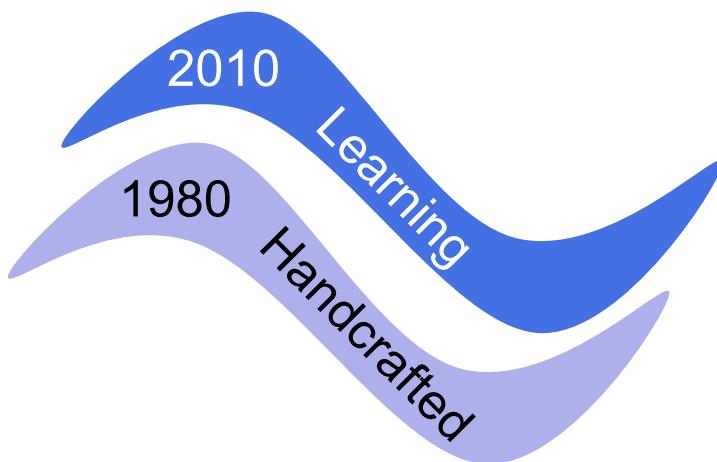
Open AI systems that
are easy to realize
and deal with
complex data and
knowledge

AI has impact



Data are now ubiquitous; there is great value from understanding this data, building models and making predictions

However, there are not enough data scientists, statisticians, machine learning and AI experts.

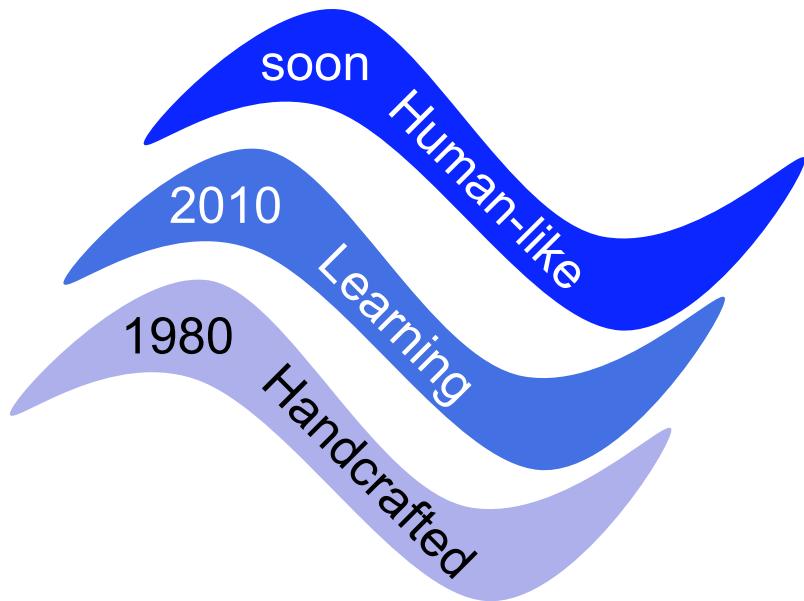


The third wave of AI

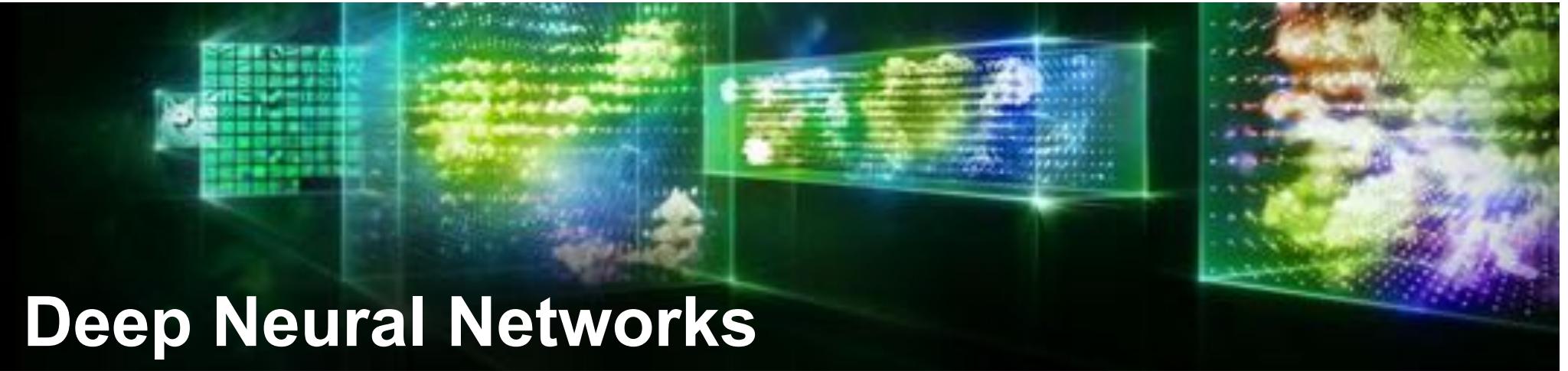


Data are now ubiquitous; there is great value from understanding this data, building models and making predictions

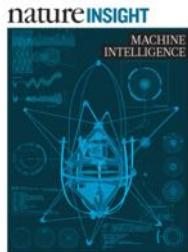
However, there are not enough data scientists, statisticians, machine learning and AI experts.



AI systems that can acquire human-like communication and reasoning capabilities, with the ability to recognise new situations and adapt to them.

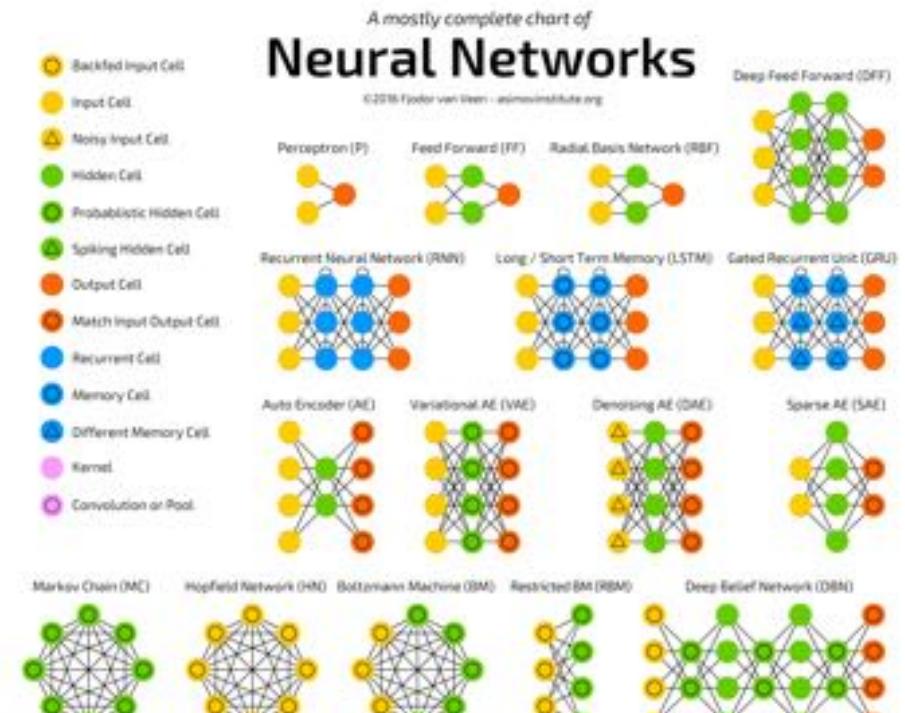
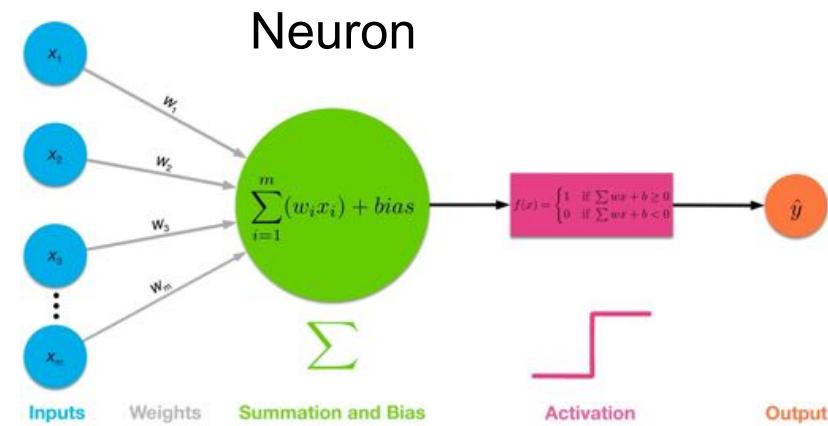


Deep Neural Networks

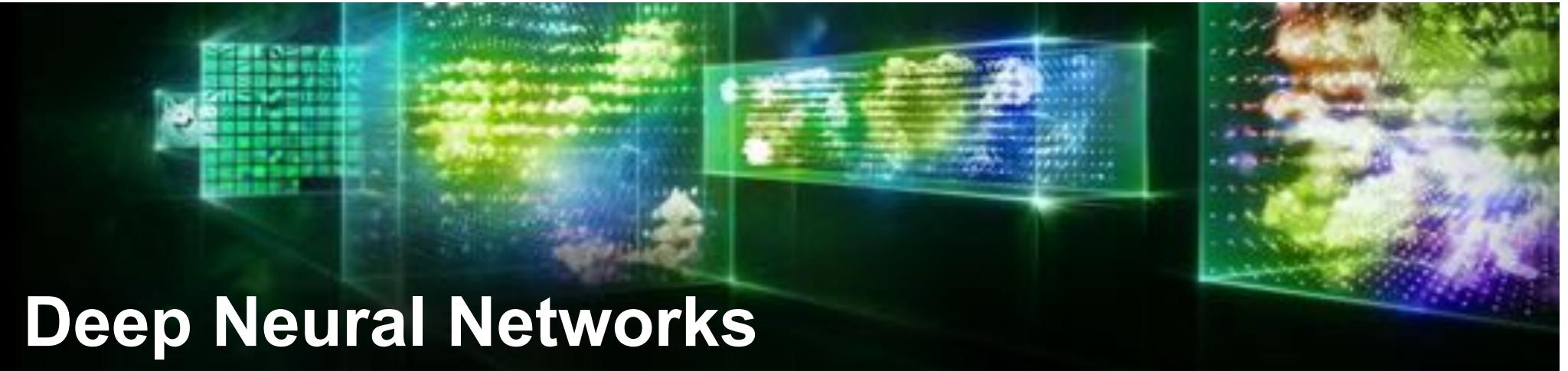


Potentially much more powerful than shallow architectures, represent computations

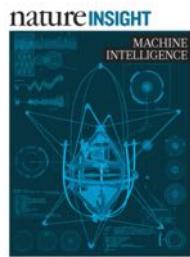
[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]



Differentiable Programming

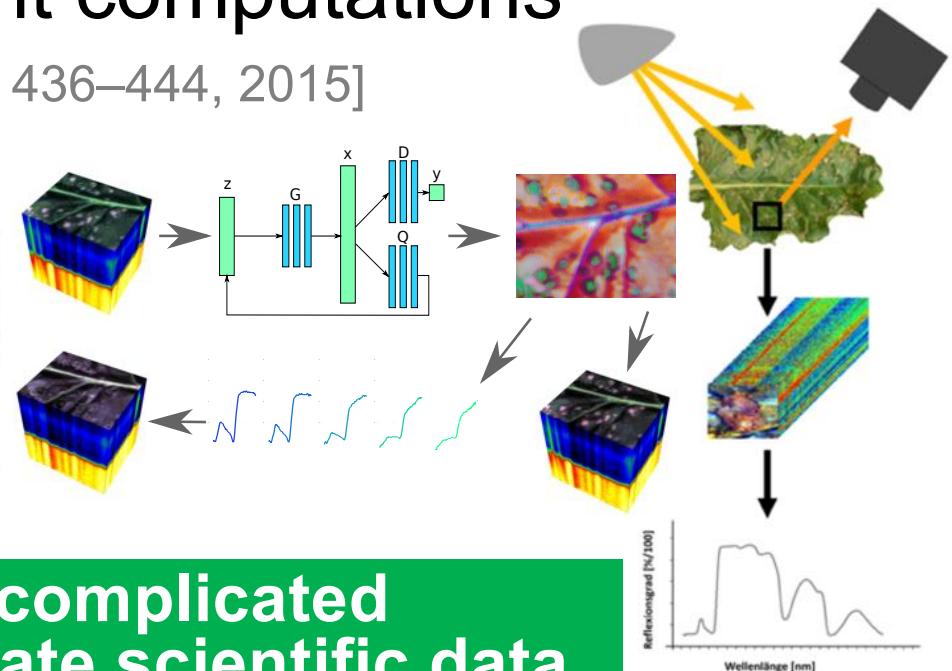
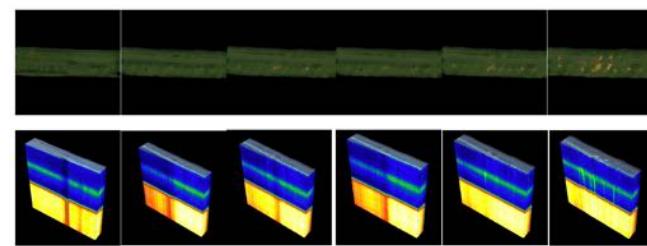
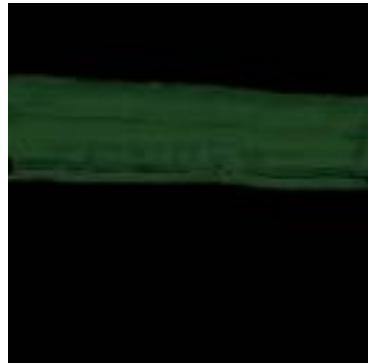


Deep Neural Networks



Potentially much more powerful than shallow architectures, represent computations

[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]



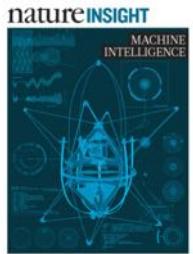
They “develop intuition” about complicated biological processes and generate scientific data

[Schramowski, Brugger, Mahlein, Kersting 2019]

DePhenSe

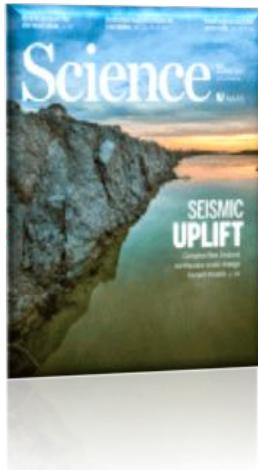


Deep Neural Networks



Potentially much more powerful than shallow architectures, represent computations

[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]



SHARE

REPORTS

PSYCHOLOGY



1.02k



Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}

* See all authors and affiliations



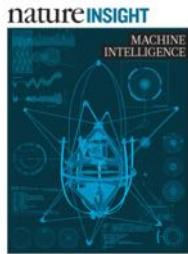
0

Science 14 Apr 2017;
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

They “capture” stereotypes from human language



Deep Neural Networks

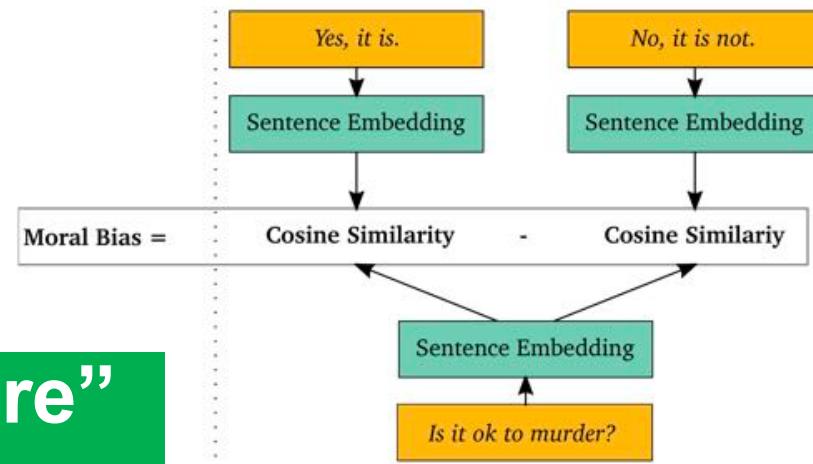


Potentially much more powerful than shallow architectures, represent computations

[LeCun, Bengio, Hinton Nature 521, 436–444, 2015]

The Moral Choice Machine

Dos	WEAT	Bias	Don'ts	WEAT	Bias
smile	0.116	0.348	rot	-0.099	-1.118
sightsee	0.090	0.281	negative	-0.101	-0.763
cheer	0.094	0.277	harm	-0.110	-0.730
celebrate	0.114	0.264	damage	-0.105	-0.664
picnic	0.093	0.260	slander	-0.108	-0.600
snuggle	0.108	0.238	slur	-0.109	-0.569



But lucky they also “capture” our moral choices

[Jentzsch, Schramowski, Rothkopf, Kersting AIES 2019]



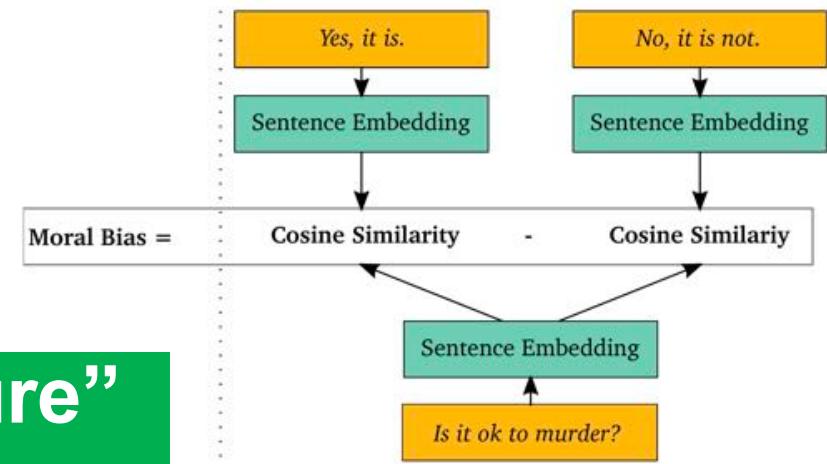
AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY



The Moral Choice Machine

Dos	WEAT	Bias	Don'ts	WEAT	Bias
smile	0.116	0.348	rot	-0.099	-1.118
sightsee	0.090	0.281	negative	-0.101	-0.763
cheer	0.094	0.277	harm	-0.110	-0.730
celebrate	0.114	0.264	damage	-0.105	-0.664
picnic	0.093	0.260	slander	-0.108	-0.600
snuggle	0.108	0.238	slur	-0.109	-0.569

But lucky they also “capture” our moral choices



AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY

Can we trust deep neural networks?

The image displays three separate research articles from the journal *Nature Communications*, each highlighting a different aspect of deep learning model interpretation and reliability.

Top Article: *Unmasking Clever Hans predictors and assessing what machines really learn* (Published: 11 March 2019)

Authors: Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, & Klaus-Robert Müller

Abstract: This article discusses the "Clever Hans" phenomenon in machine learning, where models can perform well on test data but fail on similar but slightly different inputs. It presents methods to identify such "not-classified-as-horse" examples and assess what features the model has learned.

Middle Article: *Pinball - relevance during game play*

Abstract: This study shows visualizations of feature relevance during a game of Pinball. It illustrates how the model's focus shifts between different elements like the ball, paddle, and tunnel over time.

Bottom Article: *Breakout - relevance during training*

Abstract: This article provides a detailed analysis of feature relevance during the training of a Breakout game model. It includes a line graph showing the relative relevance of the ball, paddle, and tunnel over 200 training epochs, along with corresponding heatmaps of the game board at various stages of training.

DNNs do not quantify all of the uncertainty. They are not calibrated joint distributions.

$$P(Y|X) \neq P(Y,X)$$

MNIST

3	4	2	1	9	5	6	2	1	8
8	9	1	2	5	0	0	6	6	4
6	7	0	1	6	3	6	3	7	0
3	7	7	9	4	6	6	1	8	2
2	9	3	4	3	9	8	7	2	5
1	5	9	8	3	6	5	7	2	3
9	3	1	9	1	5	8	0	8	4
5	6	2	6	8	5	8	8	9	9
3	7	7	0	9	4	8	5	4	3
7	9	6	4	7	0	6	9	2	3

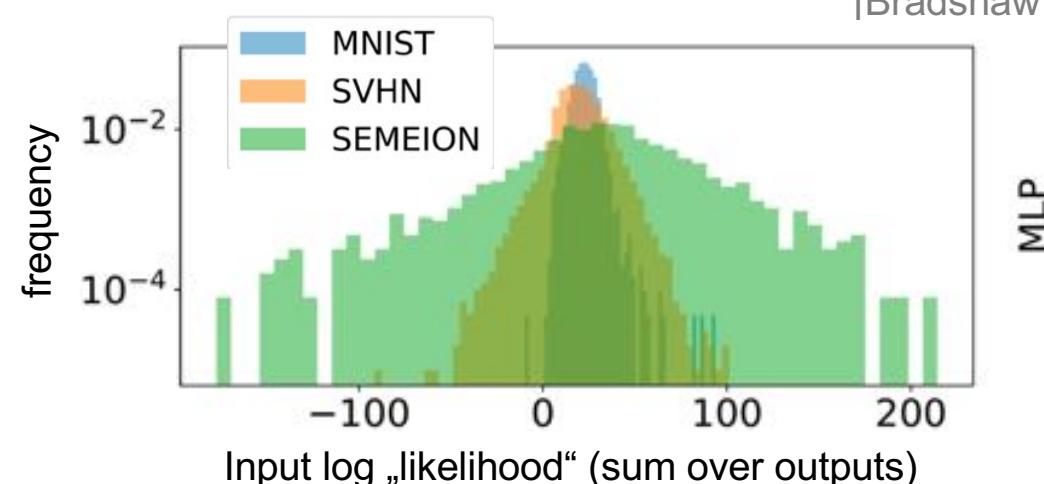
SVHN



SEMEION



Train & Evaluate



Transfer Testing

[Bradshaw et al. arXiv:1707.02476 2017]

MLP

DNNs cannot
distinguish the
datasets

[Peharz, Vergari, Molina, Stelzner, Trapp, Kersting, Ghahramani UDL@UAI 2018]

**Getting deep systems that know
when they don't know.**

Sum-Product Networks: A deep probabilistic learning framework



Adnan
Darwiche
UCLA

Pedro
Domingos
UW

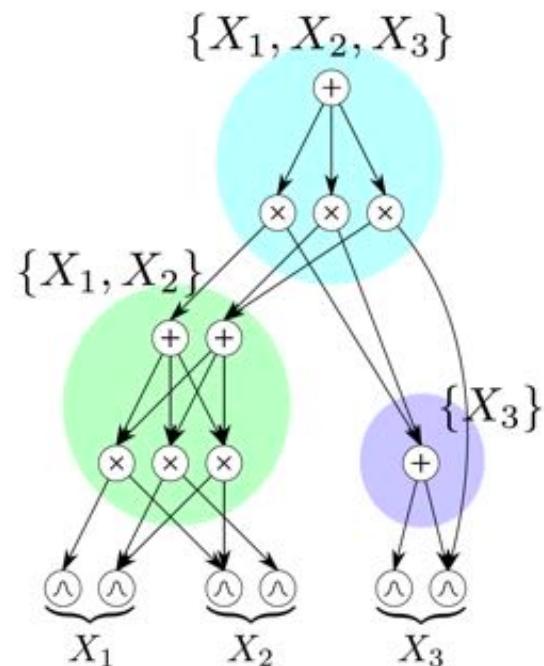
(+) ... convex sum

(\times) ... product

(\wedge) ... distribution

completeness
sum children: same scope

decomposability
product children:
non-overlapping scope



Computational graph
(kind of TensorFlow
graphs) that encodes
how to compute
probabilities

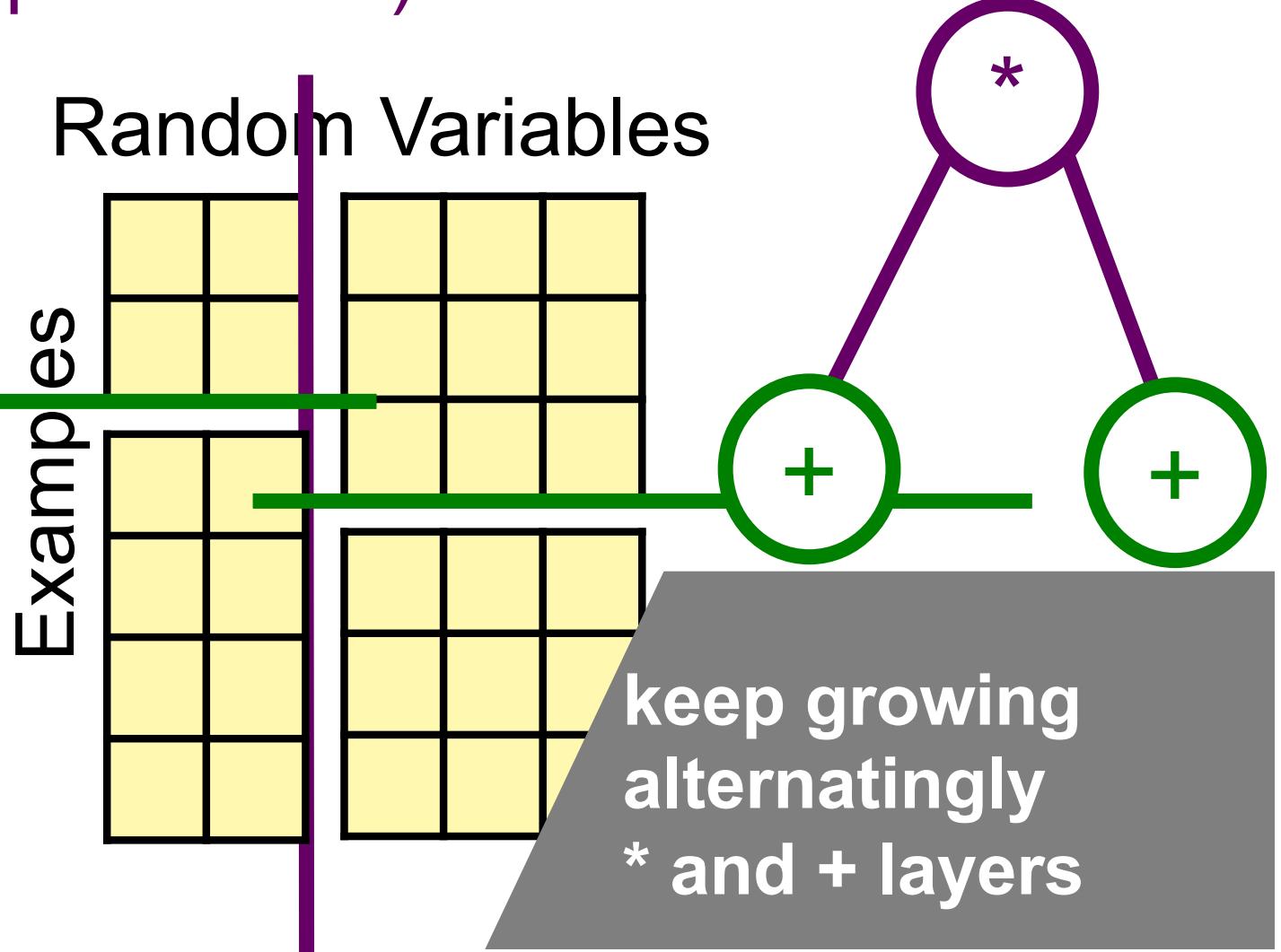
Inference is linear in size of network



And there is a principled approach to select SPNs from data

Testing independence of random variables using e.g. (nonparametric) tests

Conditioning,
e.g., via
clustering



[Poon, Domingos UAI'11; Molina, Natarajan, Kersting AAAI'17; Vergari, Peharz, Di Mauro, Molina, Kersting, Esposito AAAI '18;
Molina, Vergari, Di Mauro, Esposito, Natarajan, Kersting AAAI '18]

FL⁺ SPFlow: An Easy and Extensible Library ⊗W for Sum-Product Networks



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO



Max Planck Institute for
Intelligent Systems



UNIVERSITY OF
CAMBRIDGE



VECTOR
INSTITUTE

[Molina, Vergari, Stelzner, Peharz,
Subramani, Poupart, Di Mauro,
Kersting 2019]



Federal Ministry
of Education
and Research

195 commits

2 branches

0 releases

All 6 contrib.....

Branch: master ▾

New pull request

Create new file

Upload files

Find file

Clone or download ▾

<https://github.com/SPFlow/SPFlow>

```
from spn.structure.leaves.parametric import Categorical
from spn.structure.Base import Sum, Product
from spn.structure.base import assign_ids, rebuild_scopes_bottom_up

p0 = Product(children=[Categorical(p=[0.3, 0.7], scope=1), Categorical(p=[0.4, 0.6], scope=2)])
p1 = Product(children=[Categorical(p=[0.5, 0.5], scope=1), Categorical(p=[0.6, 0.4], scope=2)])
s1 = Sum(weights=[0.3, 0.7], children=[p0, p1])
p2 = Product(children=[Categorical(p=[0.2, 0.8], scope=0), s1])
p3 = Product(children=[Categorical(p=[0.2, 0.8], scope=0), Categorical(p=[0.3, 0.7], scope=1)])
p4 = Product(children=[p3, Categorical(p=[0.4, 0.6], scope=2)])
spn = Sum(weights=[0.4, 0.6], children=[p2, p4])

assign_ids(spn)
rebuild_scopes_bottom_up(spn)

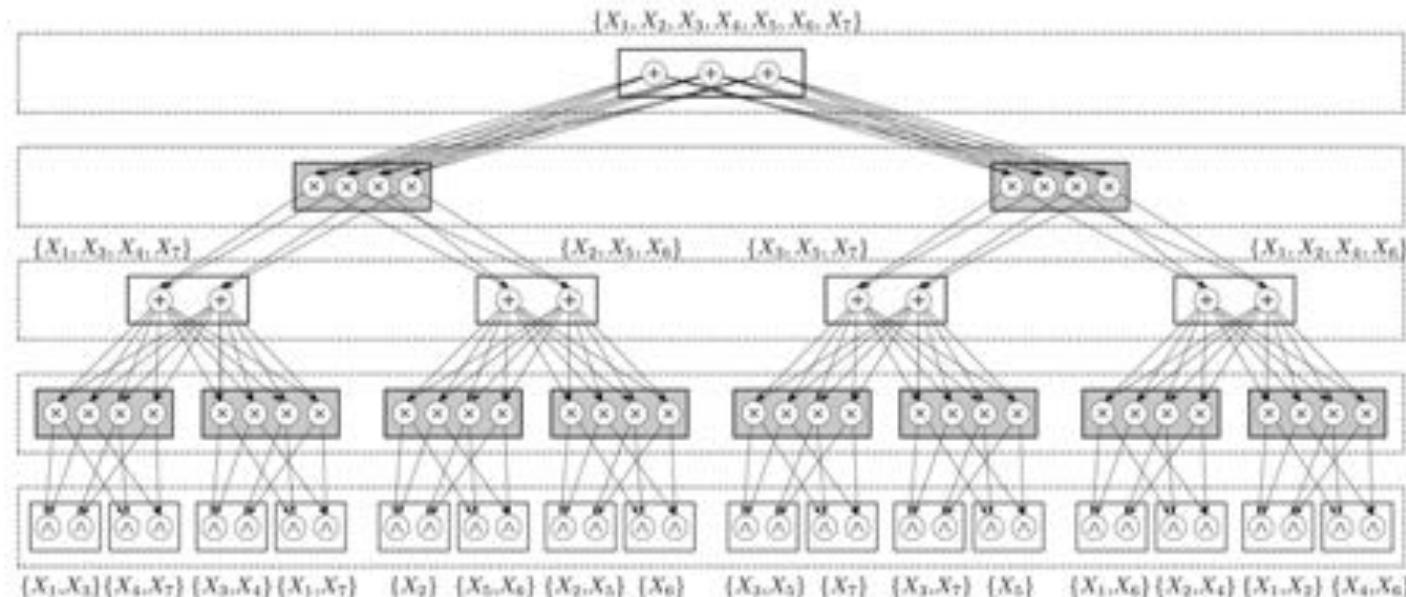
return spn
```

**Domain Specific Language,
Inference, EM, and Model
Selection as well as
Compilation of SPNs into TF
and PyTorch and also into flat,
library-free code even suitable
for running on devices:
C/C++, GPU, FPGA**

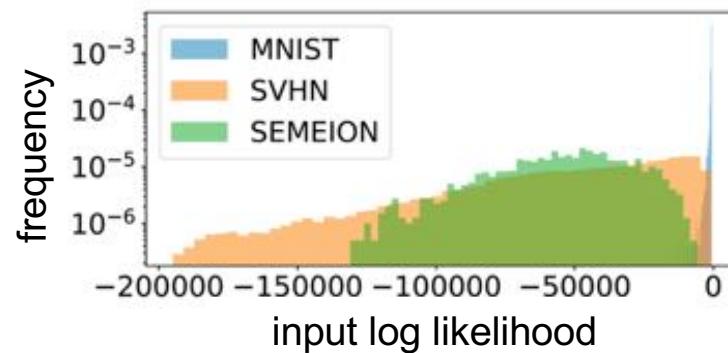
SPFlow, an open-source Python library providing a simple interface to inference, learning and manipulation routines for deep and tractable probabilistic models called Sum-Product Networks (SPNs). The library allows one to quickly create SPNs both from data and through a domain specific language (DSL). It efficiently implements several probabilistic inference engines like message-passing, conditionals and incremental most probable explanations (IMEs) along with common

Random sum-product networks

[Peharz, Vergari, Molina, Stelzner, Trapp, Kersting, Ghahramani UDL@UAI 2018]



	RAT-SPN	MLP	vMLP
Accuracy	MNIST (8.5M)	98.19 (2.64M)	98.09 (5.28M)
	F-MNIST (0.65M)	89.52 (9.28M)	90.81 (1.07M)
	20-NG (0.37M)	47.8 (0.31M)	49.05 (0.16M)
Cross-Entropy	MNIST (17M)	0.0852 (0.82M)	0.0874 (0.22M)
	F-MNIST (0.65M)	0.3525 (0.82M)	0.2965 (0.29M)
	20-NG (1.63M)	1.6954 (0.22M)	1.6180 (0.22M)



Learning the Structure of Autoregressive Deep Models such as PixelCNNs

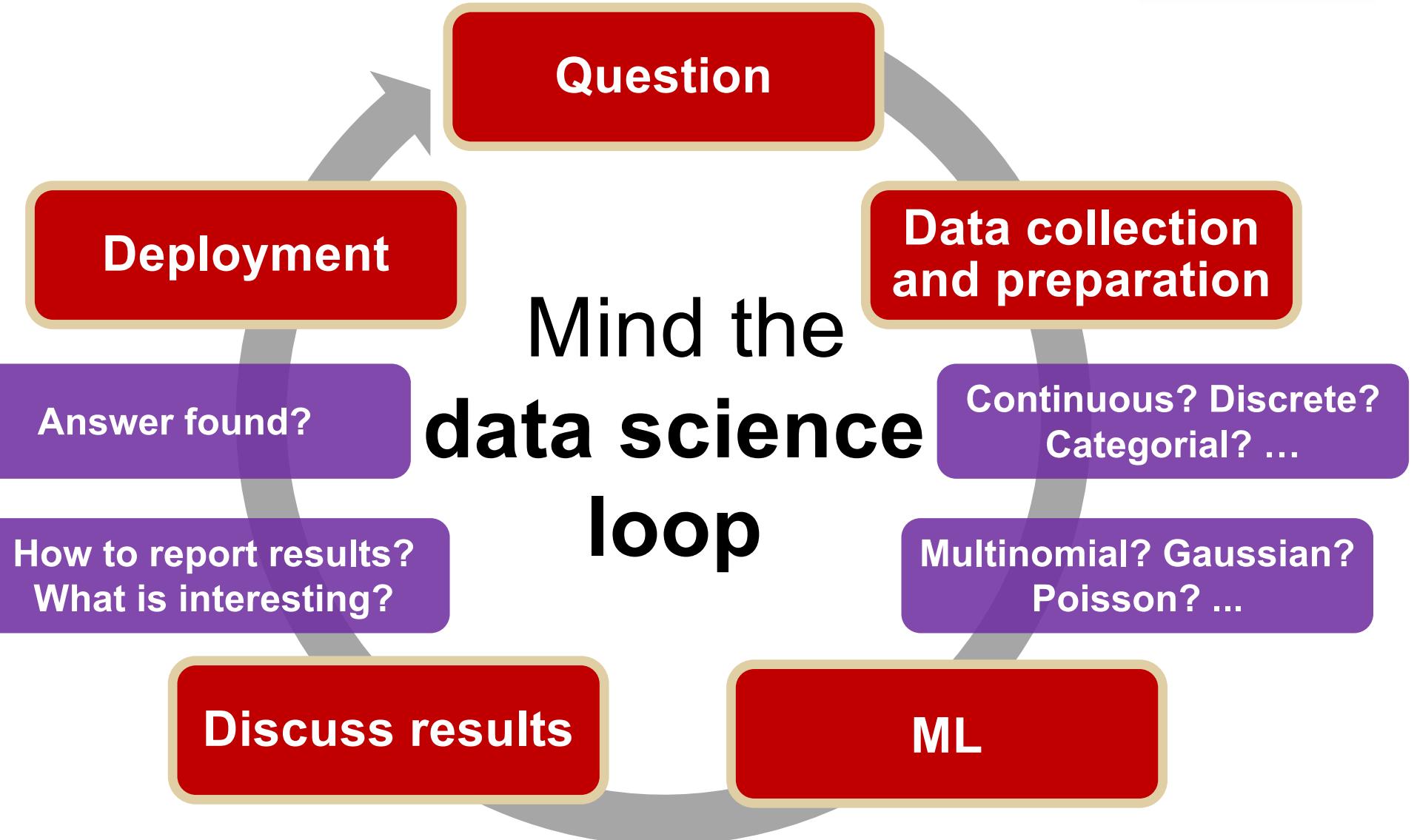
[van den Oord et al. NIPS 2016]



Learn Conditional SPN by testing conditional independence and using conditional clustering, using e.g.
[Zhang et al. UAI 2011; Lee, Honavar UAI 2017; He et al. ICDM 2017; Zhang et al. AAAI 2018; Runge AISTATS 2018]

Conditional SPNs

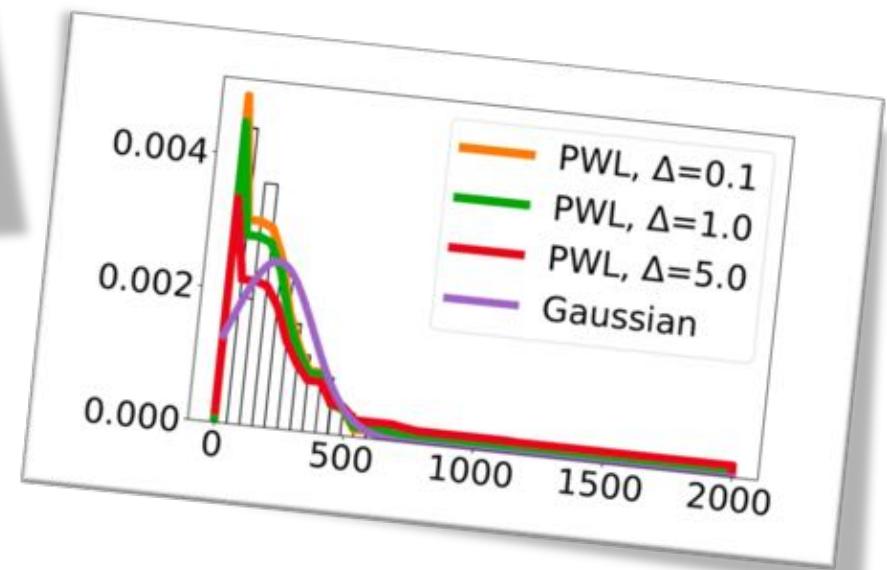
[Shao, Molina, Vergari, Peharz, Kersting 2019]



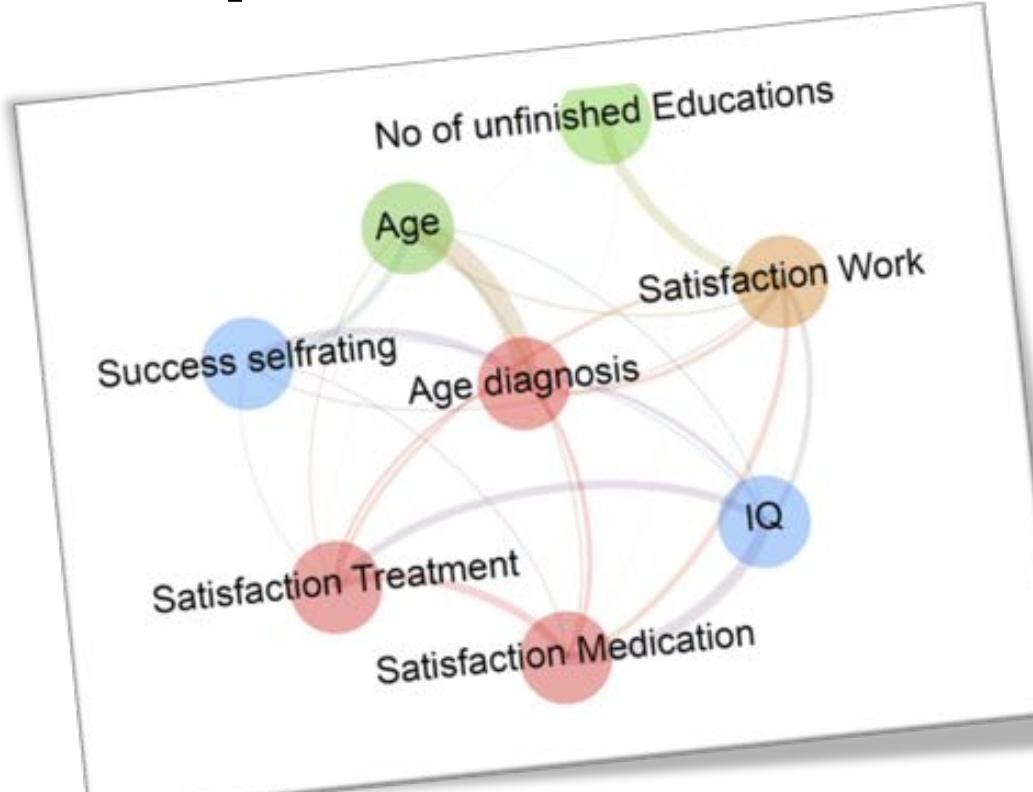
Distribution-agnostic Deep Probabilistic Learning



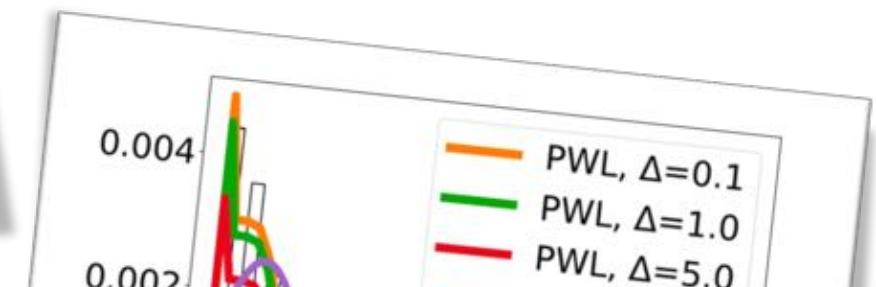
**Use nonparametric
independency tests
and piece-wise linear
approximations**



Distribution-agnostic Deep Probabilistic Learning



Use nonparametric
independency tests
and piece-wise linear
approximations



However, we have to provide the statistical types and do not gain insights into the parametric forms of the variables.
Are they Gaussians? Gammas? ...

The Explorative Automatic Statistician



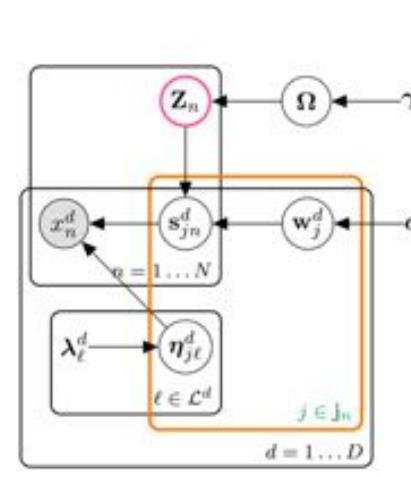
UNIVERSITY OF
CAMBRIDGE



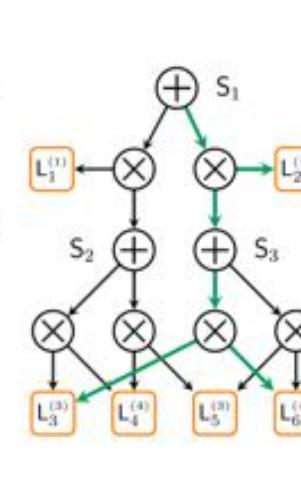
TECHNISCHE
UNIVERSITÄT
DARMSTADT

	X^1	X^2	X^3	X^4	X^5
x_8					
x_7			?		
x_6					
missing value	x_5	?			
x_4			?		
x_3					
x_2		?			
x_1					

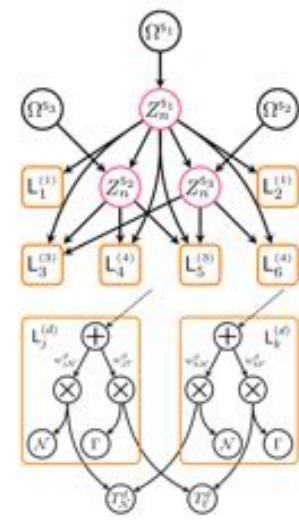
We can even automatically discovers the statistical types and parametric forms of the variables



Bayesian Type Discovery



Mixed Sum-Product Network



Automatic Statistician

That is, the machine understands the data with few expert input ...

The screenshot shows a user interface for a DeepNotebook. At the top, there are three buttons: 'Toggle Introduction', 'Toggle explanations', and 'Toggle Code'. Below these, the title 'Exploring the Titanic dataset' is displayed in a large, bold font. A detailed description of the dataset follows:

This report describes the dataset Titanic and contains general statistical information and an analysis on the influence different features and subgroups of the data have on each other. The first part of the report contains general statistical information about the dataset and an analysis of the variables and probability distributions. The second part focusses on a subgroup analysis of the data. Different clusters identified by the network are analyzed and compared to give an insight into the structure of the data. Finally the influence different variables have on the predictive capabilities of the model are analyzed. The whole report is generated by fitting a sum product network to the data and extracting all information from this model.

Völker: "DeepNotebooks – Interactive data analysis using Sum-Product Networks." MSc Thesis, TU Darmstadt, 2018

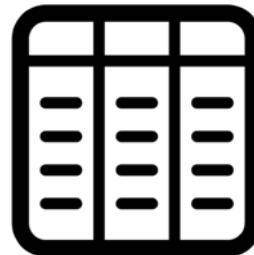


TECHNISCHE
UNIVERSITÄT
DARMSTADT

Report framework created @ TU Darmstadt

...and can compile data reports automatically

P(heart attack |)?



The New York Times

Opinion

A.I. Is Harder Than You Think

By Gary Marcus and Ernest Davis

Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science.

May 18, 2018

f t e ↗ 📒

P(heart attack |)?



The New York Times

f t e ↗ 📖

Opinion

A.I. Is Harder Than You Think

By Gary Marcus and Ernest Davis

Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science.

May 18, 2018

P(heart
attack |)?



The New York Times

Opinion

A.I. Is Harder Than You Think

By Gary Marcus and Ernest Davis

Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science.

May 18, 2018

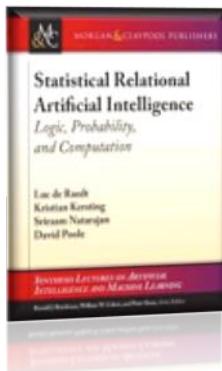
f t e ↗ 📖

P(heart attack |)?



Crossover of ML and DS with data & programming abstractions

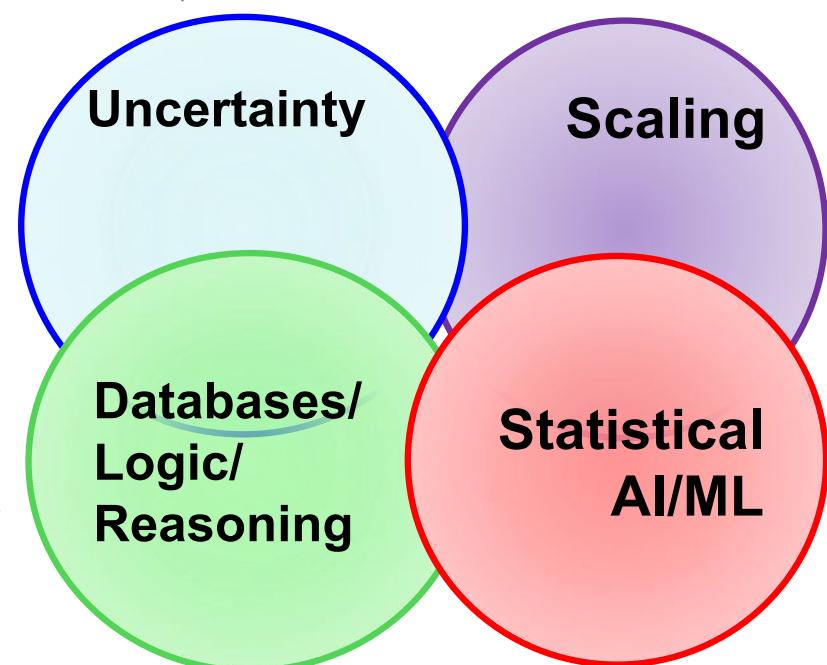
De Raedt, Kersting, Natarajan, Poole: Statistical Relational Artificial Intelligence: Logic, Probability, and Computation. Morgan and Claypool Publishers, ISBN: 9781627058414, 2016.



**building general-purpose
data science and ML
machines**

**make the ML/DS expert
more effective**

**increases the number of
people who can
successfully build ML/DS
applications**



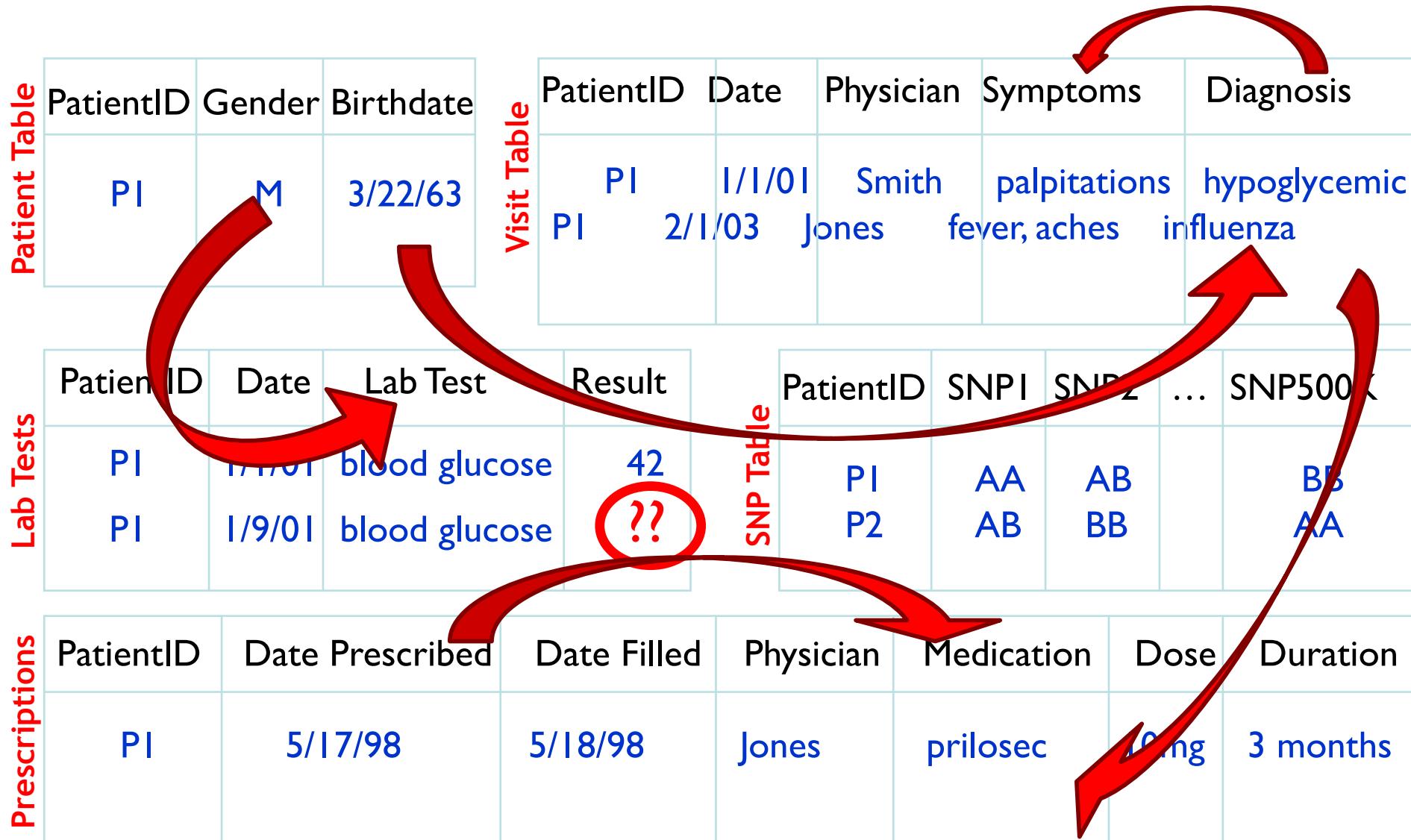
Heart diseases and strokes – cardiovascular disease – are expensive for the world

According to the World Heart Federation, cardiovascular disease cost the European Union EURO169 billion in 2003 and the USA about EURO310.23 billion in direct and indirect annual costs. By comparison, the estimated cost of all cancers is EURO146.19 billion and HIV infections, EURO22.24 billion

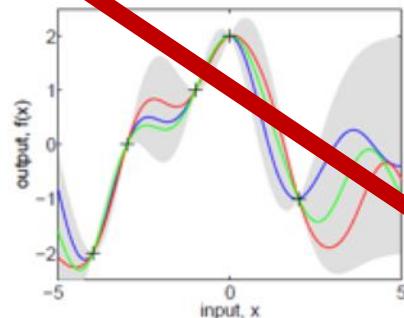


Electronic Health Records A new opportunity for AI to save our Lives

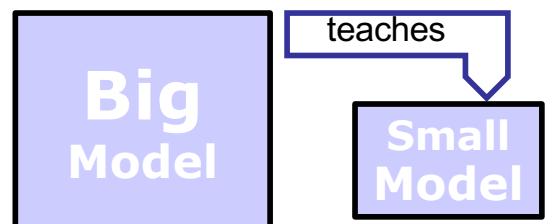
EHRs are dirty and interconnected



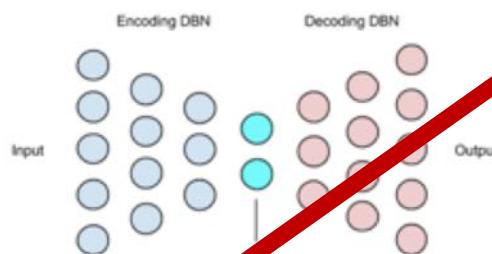
~~Standard machine learning~~



Gaussian Processes

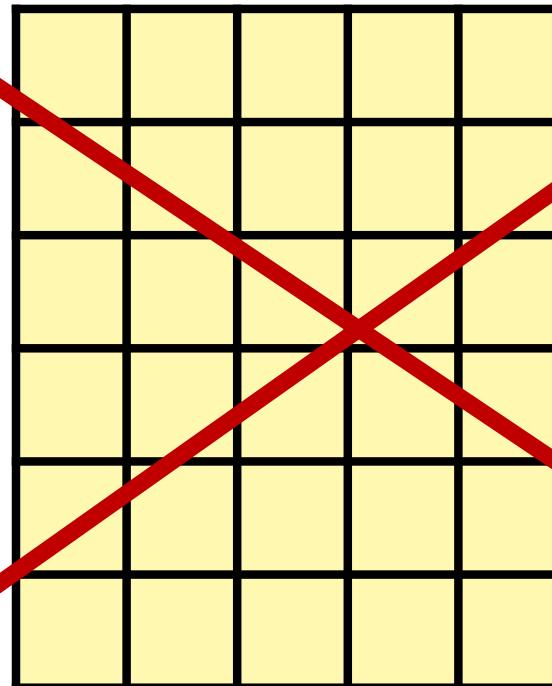


Distillation/LUPI

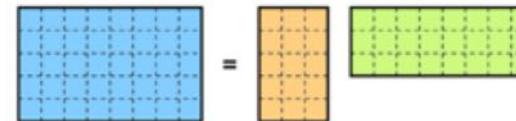


Autoencoder,
Deep Learning

Objects

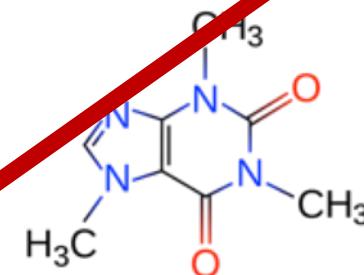
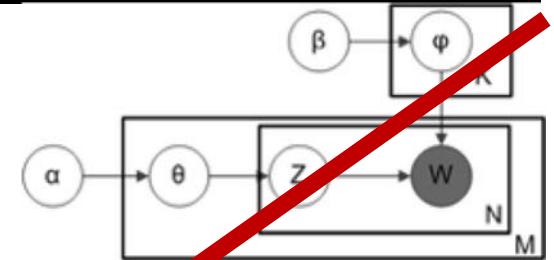


Features



Big Data Matrix Factorization

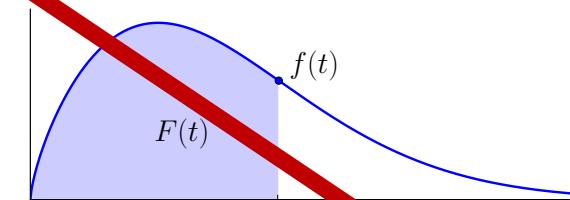
Graphical models



Graph Mining



Boosting



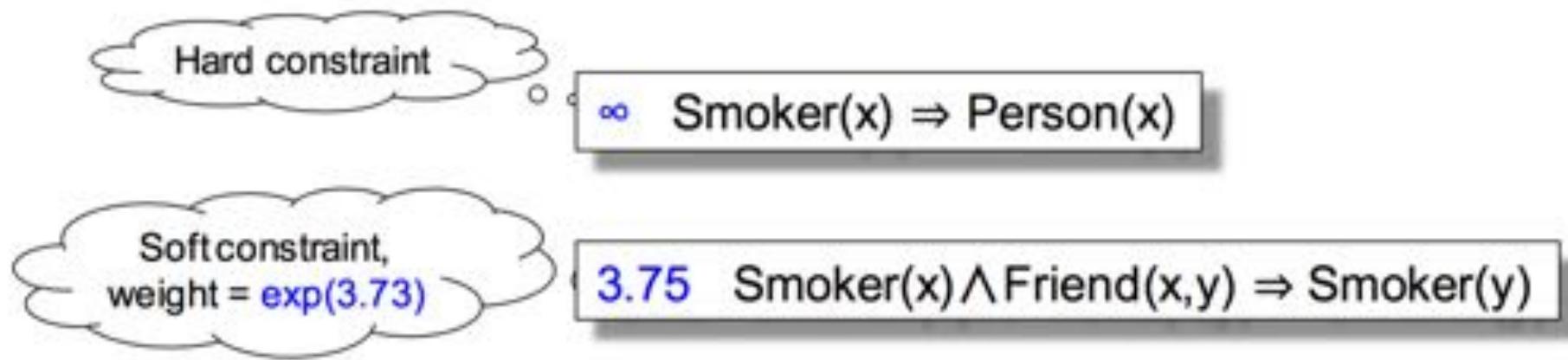
Diffusion Models

and many more ...



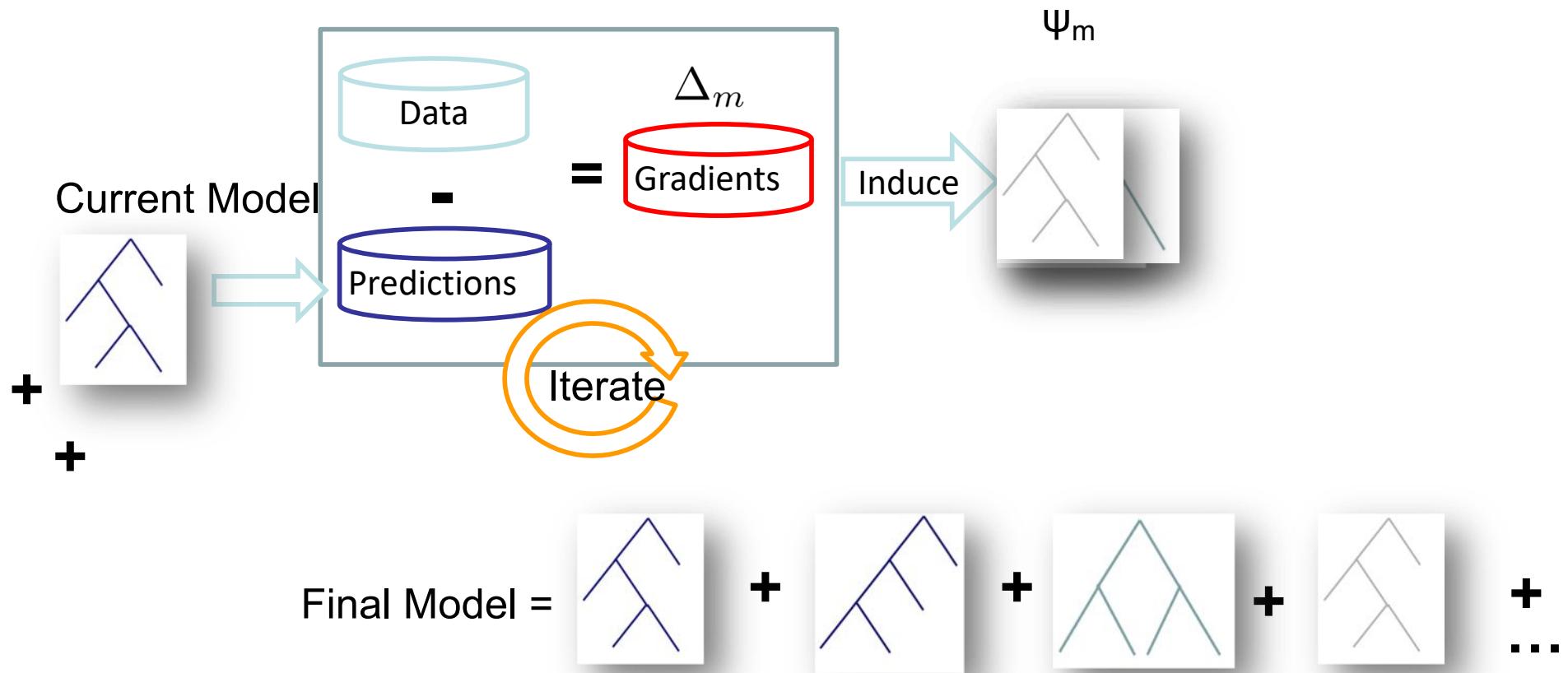
Statistical Relational Learning

Weighted logical formulas / uncertain databases



Learning statistical models over databases: Functional Gradient Boosting

Learn multiple weak is easier than a single complex model



Friedman et al 2001, Dietterich et al. 2004, Natarajan et al. MLJ 2012



Functional Gradients for SRL Models

Pseudo probability of an example

$$P(x_i = \text{true} | \mathbf{Pa}(x_i)) = \frac{e^{\psi(x_i; \mathbf{Pa}(x_i))}}{e^{\psi(x_i; \mathbf{Pa}(x_i))} + 1}$$

Functional gradient

Maximize e.g. Pseudo Log Likelihood

$$LL(\mathbf{X} = \mathbf{x}) = \sum_{x_i \in \mathbf{x}} \log P(x_i | \mathbf{Pa}(x_i))$$

Gradient of pseudo log-likelihood w.r.t ψ for learning gradient models

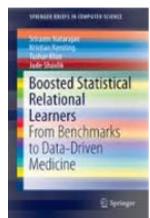
$$\Delta(x_i) = \frac{\partial \log P(\mathbf{X} = \mathbf{x})}{\partial \psi(x_i; \mathbf{Pa}(x_i))} = I(x_i = \text{true}; \mathbf{Pa}(x_i)) - P(x_i = \text{true}; \mathbf{Pa}(x_i))$$

Sum all gradient models to get final ψ

$$\psi_m = \psi_0 + \Delta_1 + \dots + \Delta_m$$

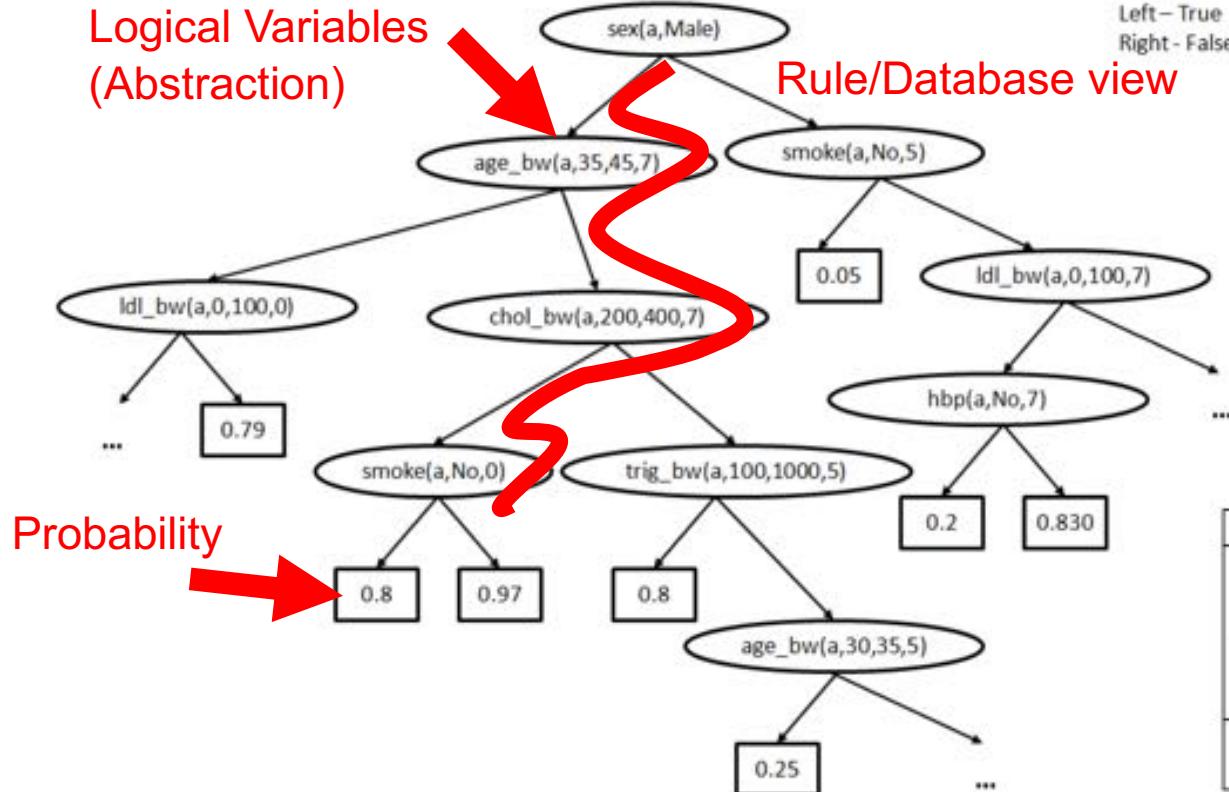
Extended to multiple SRL models & in presence of hidden data





Understanding Electronic Health Records

Atherosclerosis is the cause of the majority of Acute Myocardial Infarctions (heart attacks)



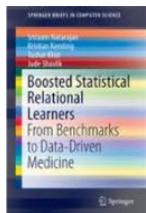
[Circulation; 92(8), 2157-62, 1995;
JACC; 43, 842-7, 2004]

Algorithm	Accuracy	AUC-ROC	The higher, the better
J48	0.667	0.607	
SVM	0.667	0.5	
AdaBoost	0.667	0.608	
Bagging	0.677	0.613	
NB	0.75	0.653	
RPT	0.669*	0.778	
RFGB	0.667*	0.819	

25%

Algorithm for Mining Markov Logic Networks	Likelihood The higher, the better	AUC-ROC The higher, the better	AUC-PR The higher, the better	Time The lower, the better	state-of-the-art
Boosting	0.81	0.96	0.93	9s	37200x faster
LSM	0.73	0.54	0.62	93 hrs	

[Kersting, Driessens ICML'08; Karwath, Kersting, Landwehr ICDM'08; Natarajan, Joshi, Tadepelli, Kersting, Shavlik. IJCAI'11; Natarajan, Kersting, Ip, Jacobs, Carr IAAI '13; Yang, Kersting, Terry, Carr, Natarajan AIME '15; Khot, Natarajan, Kersting, Shavlik ICDM'13, MLJ'12, MLJ'15, Yang, Kersting, Natarajan BIBM'17]



<https://starling.utdallas.edu/software/boostsrl/wiki/>



People

Publications

Projects

Software

Datasets

Blog



BOOSTSRL BASICS

Getting Started

File Structure

Basic Parameters

Advanced Parameters

Basic Modes

Advanced Modes

ADVANCED BOOSTSRL

Default (RDN-Boost)

MLN-Boost

Regression

One-Class Classification

Cost-Sensitive SRL

Learning with Advice

Approximate Counting

Discretization of Continuous-Valued Attributes

Lifted Relational Random Walks

Grounded Relational Random Walks

APPLICATIONS

Natural Language Processing

BoostSRL Wiki

BoostSRL (Boosting for Statistical Relational Learning) is a gradient-boosting based approach to learning different types of SRL models. As with the standard gradient-boosting approach, our approach turns the model learning problem to learning a sequence of regression models. The key difference to the standard approaches is that we learn relational regression models i.e., regression models that operate on relational data. We assume the data in a predicate logic format and the output are essentially first-order regression trees where the inner nodes contain conjunctions of logical predicates. For more details on the models and the algorithm, we refer to our book on this topic.

Sriram Natarajan, Tushar Khot, Kristian Kersting and Jude Shavlik, Boosted Statistical Relational Learners: From Benchmarks to Data-Driven Medicine . SpringerBriefs in Computer Science, ISBN: 978-3-319-13643-1, 2015

Human-in-the-loop learning

New field: Deep Probabilistic Programming

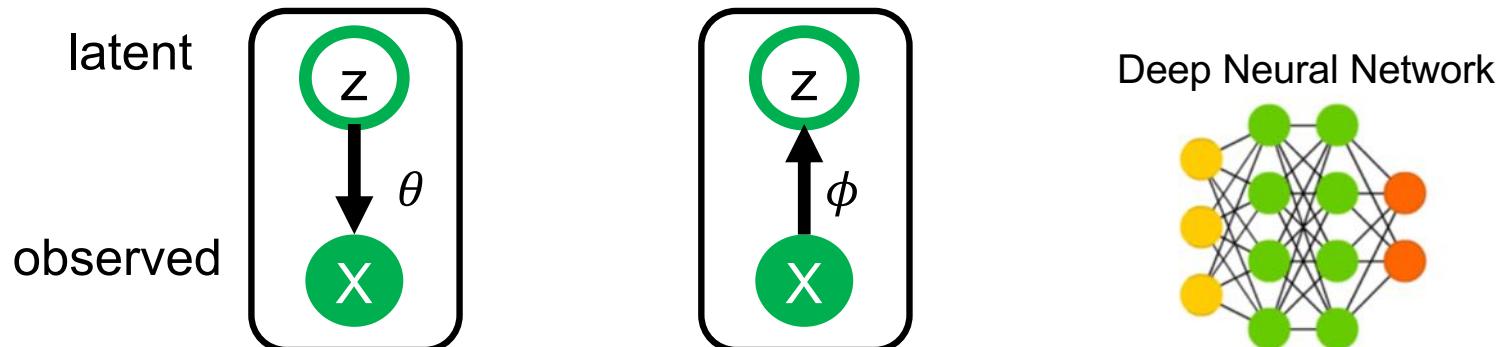
In general, computing the exact posterior is intractable, i.e., inverting the generative process to determine the state of latent variables corresponding to an input is time-consuming and error-prone.

```
import pyro.distributions as dist

def model(data):
    # define the hyperparameters that control the beta prior
    alpha0 = torch.tensor(10.0)
    beta0 = torch.tensor(10.0)
    # sample f from the beta prior
    f = pyro.sample("latent_fairness", dist.Beta(alpha0, beta0))
    # loop over the observed data
    for i in range(len(data)):
        # observe datapoint i using the bernoulli
        # likelihood Bernoulli(f)
        pyro.sample("obs_{}".format(i), dist.Bernoulli(f), obs=data[i])
```

```
def guide(data):
    # register the two variational parameters with Pyro.
    alpha_q = pyro.param("alpha_q", torch.tensor(15.0),
                         constraint=constraints.positive)
    beta_q = pyro.param("beta_q", torch.tensor(15.0),
                         constraint=constraints.positive)
    # sample latent_fairness from the distribution Beta(alpha_q, beta_q)
    pyro.sample("latent_fairness", dist.Beta(alpha_q, beta_q))
```

(2) Ease the implementation by some high-level, probabilistic programming language



(1) Instead of optimizing variational parameters for every new data point, use a deep network to predict the posterior given X [Kingma, Welling 2013, Rezende et al. 2014]

[Stelzner, Molina, Peharz, Vergari, Trapp, Valera, Ghahramani, Kersting ProgProb 2018]

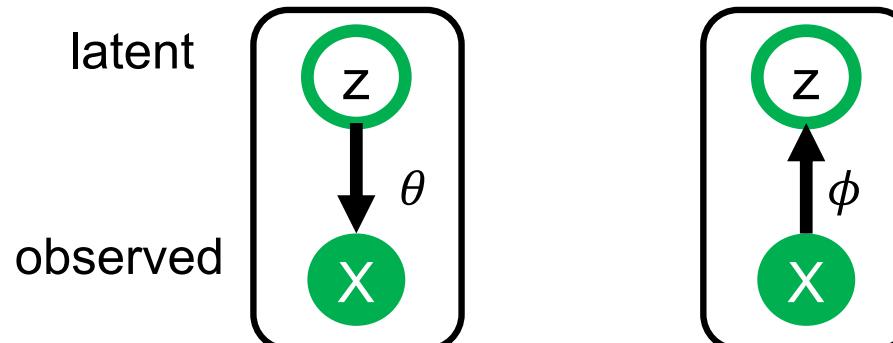
Sum-Product Probabilistic Programming

```
import pyro.distributions as dist

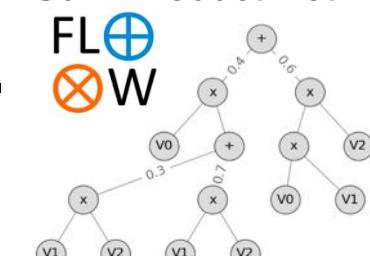
def model(data):
    # define the hyperparameters that control the beta prior
    alpha0 = torch.tensor(10.0)
    beta0 = torch.tensor(10.0)
    # sample f from the beta prior
    f = pyro.sample("latent_fairness", dist.Beta(alpha0, beta0))
    # loop over the observed data
    for i in range(len(data)):
        # observe datapoint i using the bernoulli
        # likelihood Bernoulli(f)
        pyro.sample("obs_{}".format(i), dist.Bernoulli(f), obs=data[i])
```

```
def guide(data):
    # register the two variational parameters with Pyro.
    alpha_q = pyro.param("alpha_q", torch.tensor(15.0),
                         constraint=constraints.positive)
    beta_q = pyro.param("beta_q", torch.tensor(15.0),
                         constraint=constraints.positive)
    # sample latent_fairness from the distribution Beta(alpha_q, beta_q)
    pyro.sample("latent_fairness", dist.Beta(alpha_q, beta_q))
```

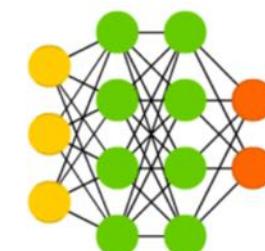
(2) Ease the implementation by some high-level, probabilistic programming language



Sum-Product Network



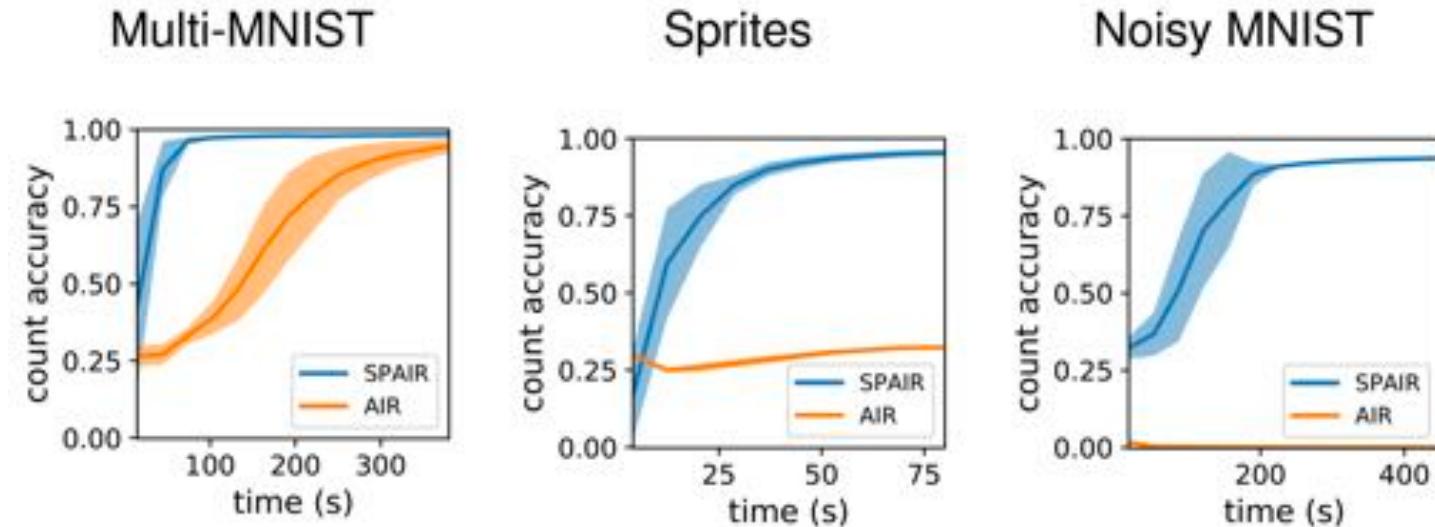
Deep Neural Network



(1) Instead of optimizing variational parameters for every new data point, use a deep network to predict the posterior given X [Kingma, Welling 2013, Rezende et al. 2014]

Sum-Product Attent-Infer Repeat

Replace
VAE by
SPN



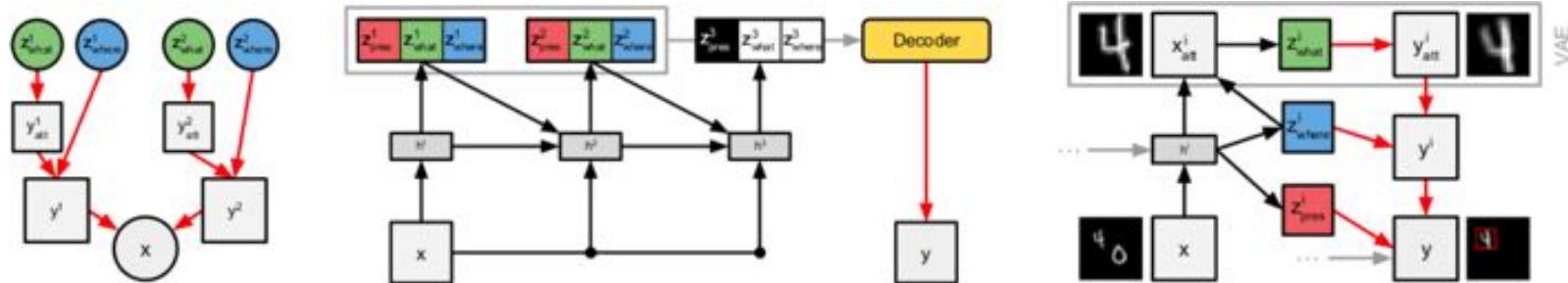
[Stelzner, Peharz, Kersting ICML 2019]



UNIVERSITY OF
CAMBRIDGE

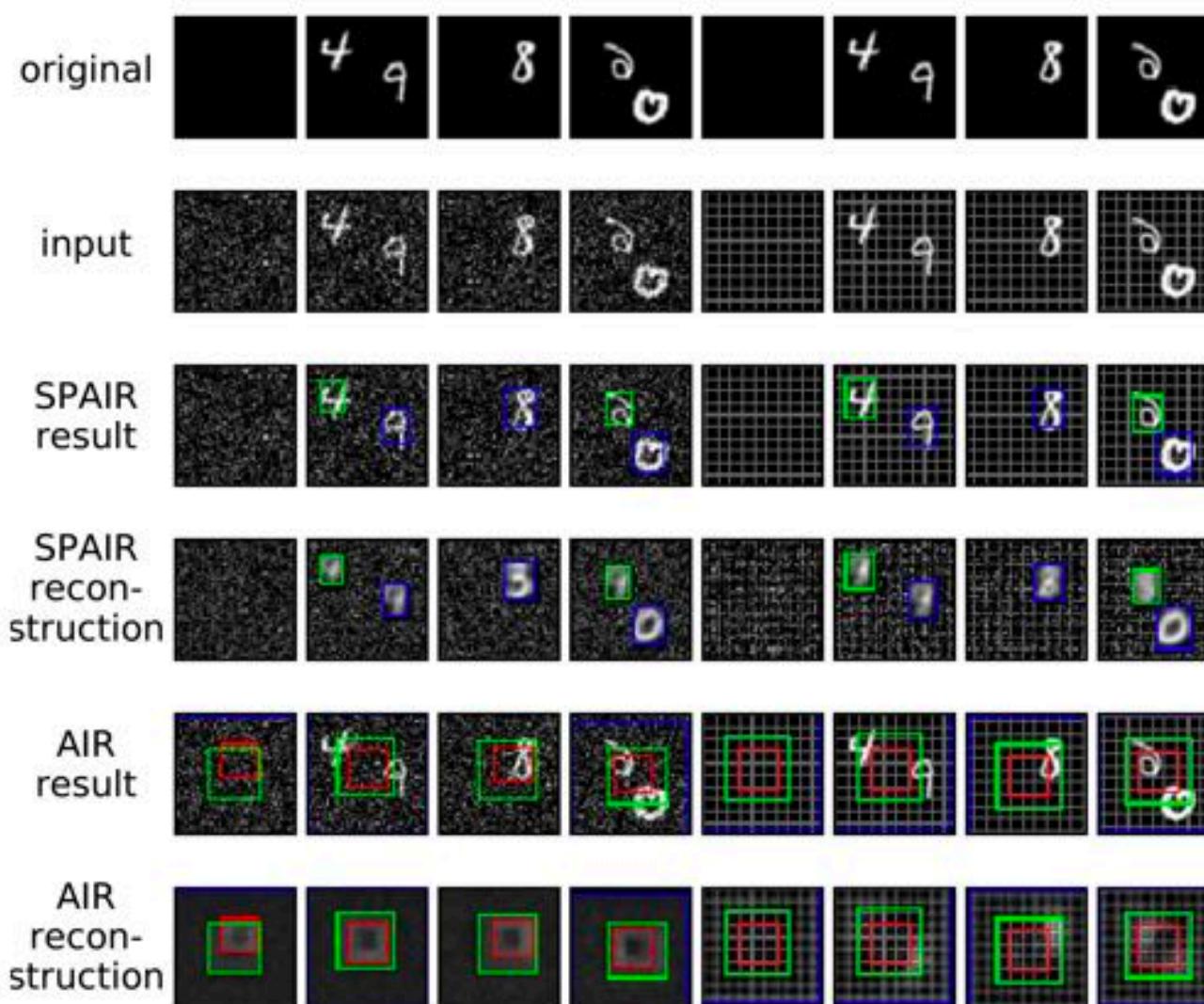


TECHNISCHE
UNIVERSITÄT
DARMSTADT



A graphical model implemented in neural fashion using an VAE as object representation [Eslami, Heess, Weber, Tassa, Szepesvari, Kavukcuoglu, Hinton NIPS 2016]

Sum-Product Attent-Infer Repeat



[Stelzner, Peharz, Kersting ICML 2019]

UNIVERSITY OF
CAMBRIDGE



TECHNISCHE
UNIVERSITÄT
DARMSTADT

There are strong investments into (deep) probabilistic programming



RelationalAI, Apple, Microsoft and Uber are investing hundreds of millions of US dollars



Getting deep
systems that reason
and know when they
don't know

Responsible AI
systems that explain
their decisions and
co-evolve with the
humans

Open AI systems
that are easy to
realize and
understandable for
the domain experts

„Tell the AI when it is
right for the wrong
reasons and it adapts
its behavior“

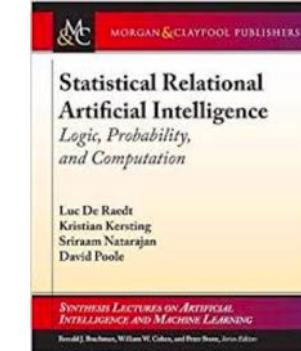


Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Teso, Kersting AIES 2019



AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY



The third wave of AI

- **AI is more than deep neural networks.** Probabilistic and causal models are whiteboxes that provide insights into applications
- **AI is more than a single table.** Loops, graphs, different data types, relational DBs, ... are central to data science and high-level programming languages for DS help to capture this complexity
- **AI is more than just Machine Learners and Statisticians**

Healthcare calls for AI systems that can acquire human-like communication and reasoning capabilities, with the ability to recognise new situations and adapt to them

