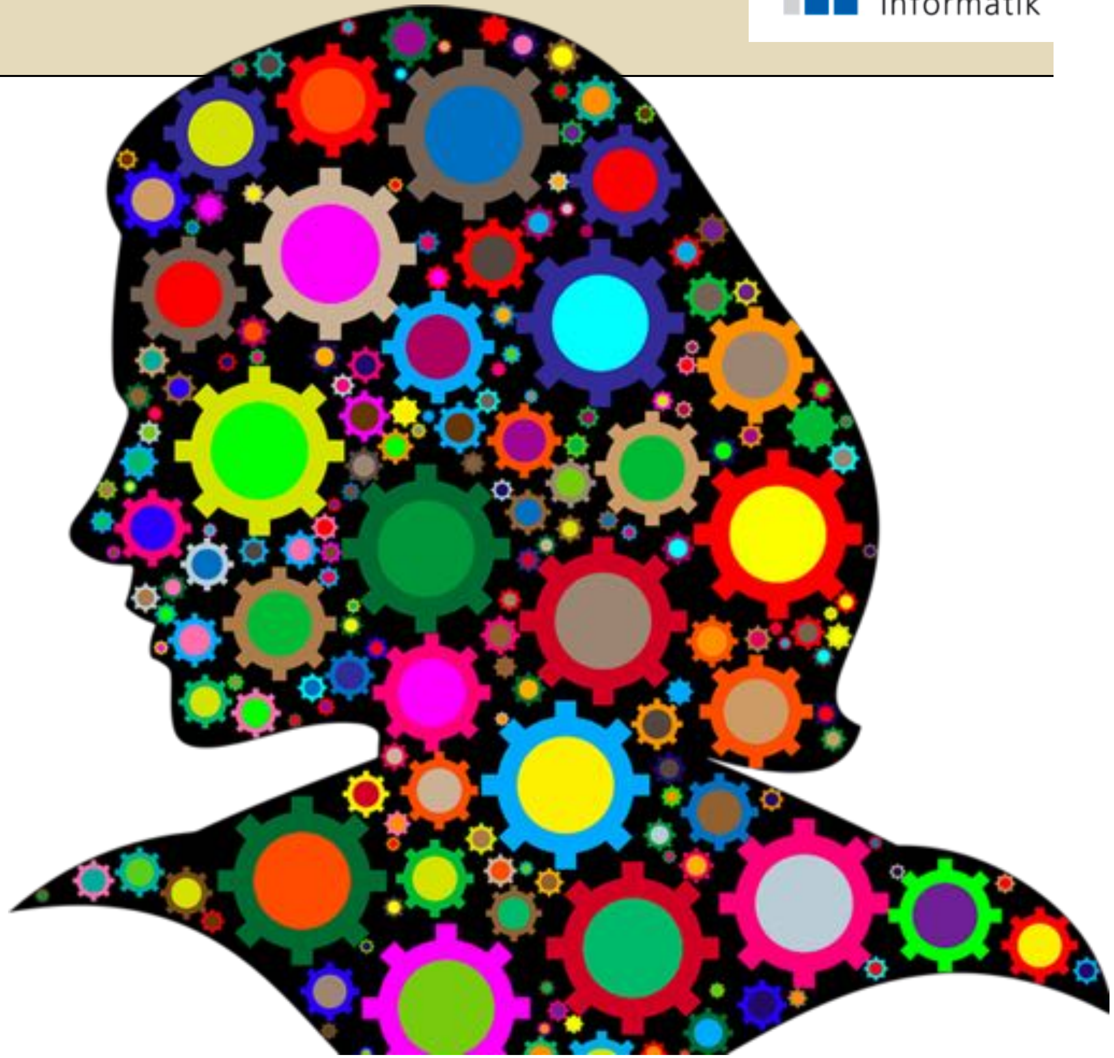
The Automatic Data Scientist

Kristian Kersting

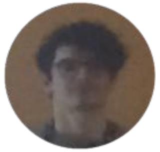
TECHNISCHE UNIVERSITÄT DARMSTADT

Centre for Cognitive Science

Fachbereich Informatik



Alejandro Molina
TU Darmstadt



Antonio Vergari
MPI-IS

Robert Peharz
U. Cambridge

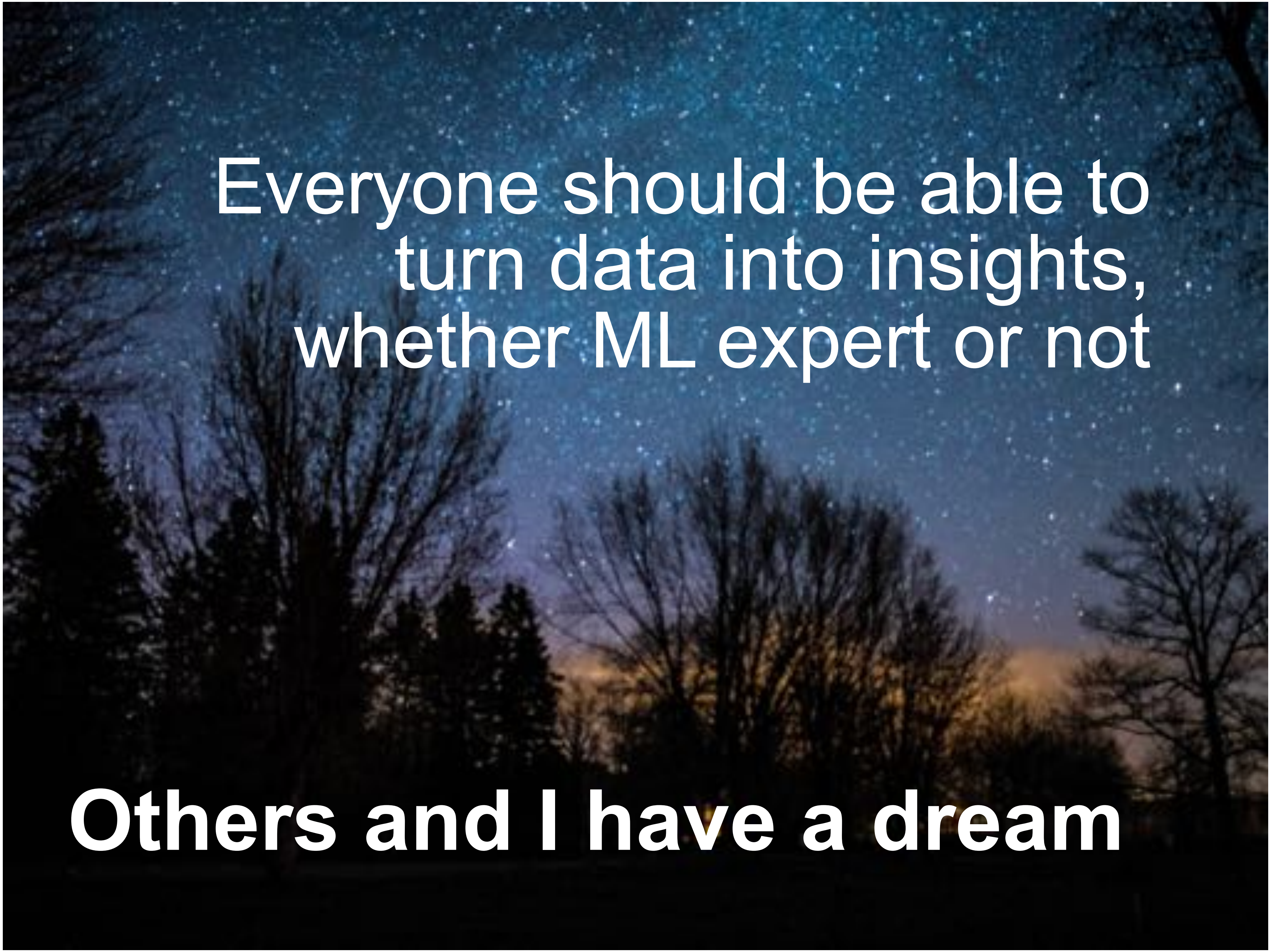
Isabell Valera
MPI-IS

Zoubin Ghahramani
UBER AI Lab,
U. Cambridge

Martin Grohe
RWTH Aachen

Martin Mladenov
Google Research

Claas Völcker
TU Darmstadt


Everyone should be able to
turn data into insights,
whether ML expert or not
Others and I have a dream

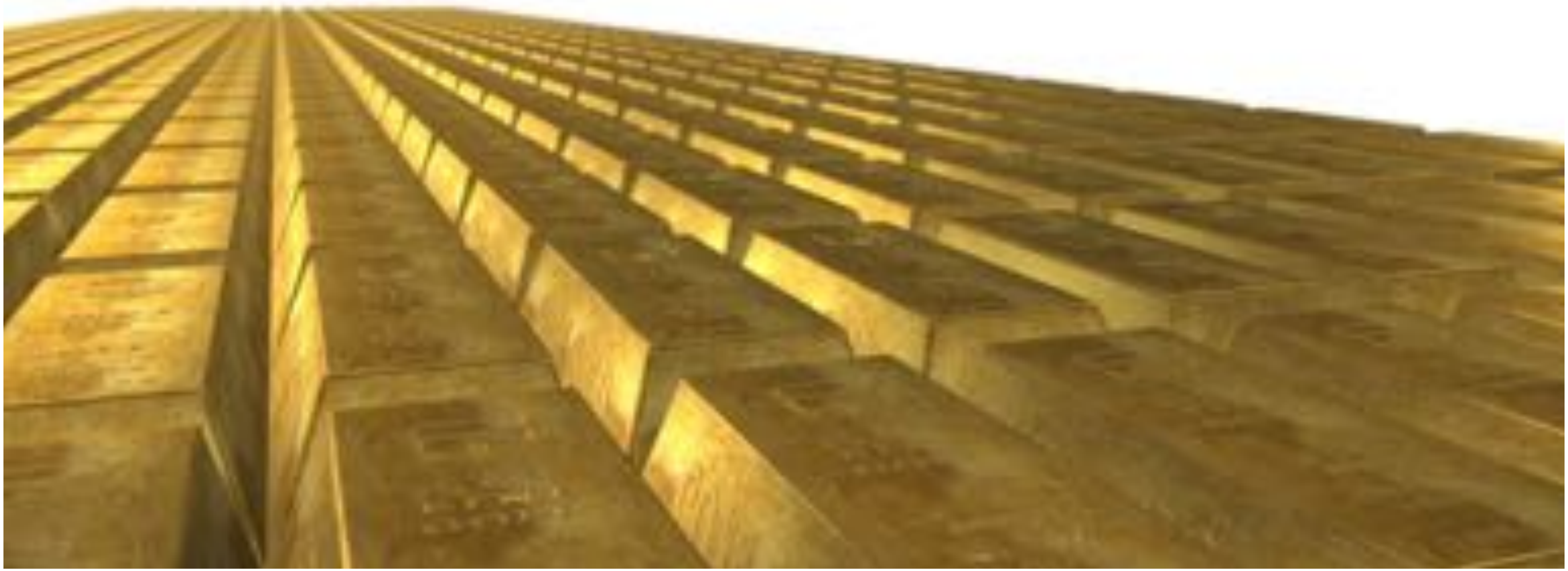
This poses many deep and fascinating questions

How can computers reason about and learn with complex data?

How can computers decide autonomously which representation is best for the data?

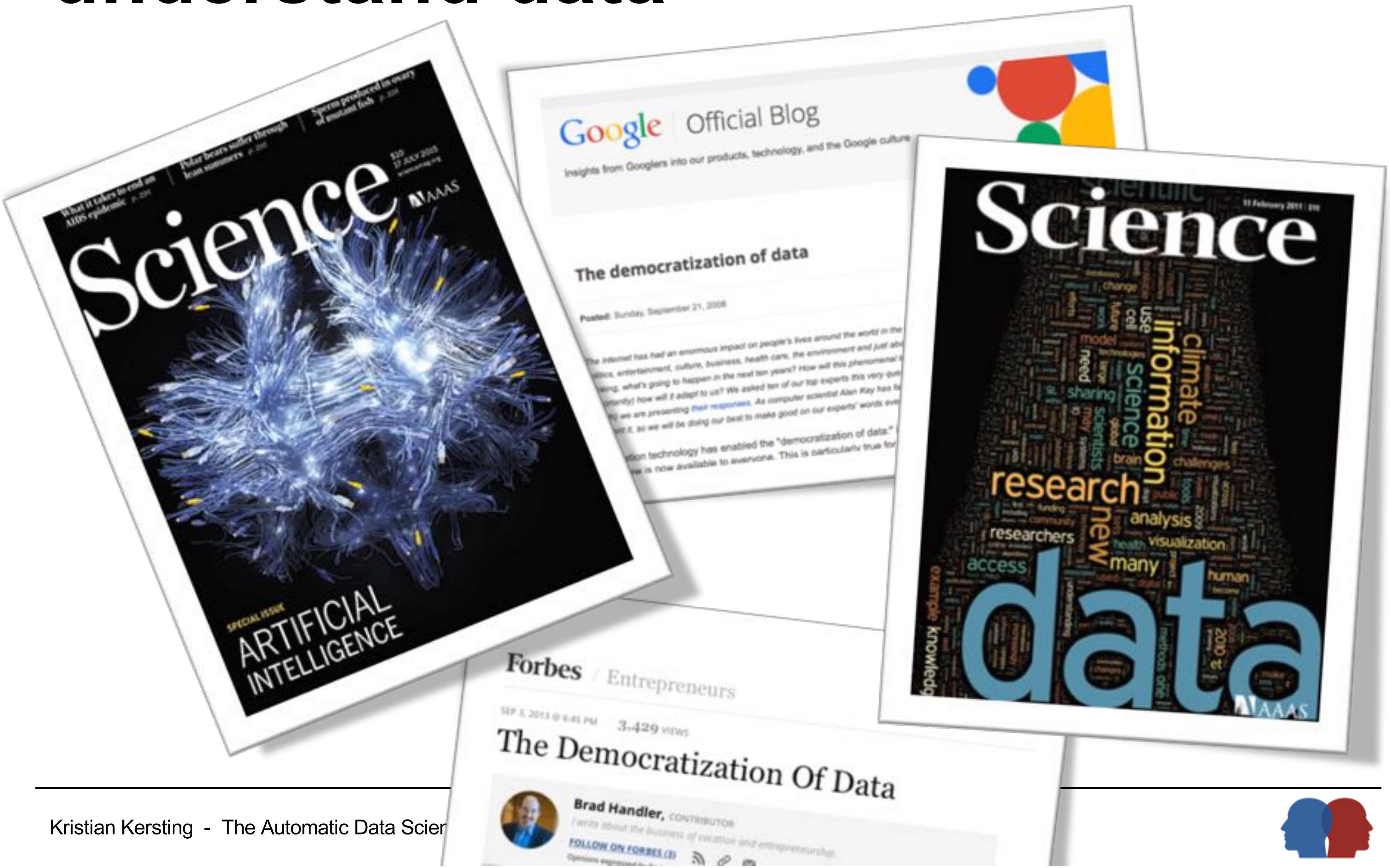
How can computers understand data with minimal expert input?

Today is the golden era of data

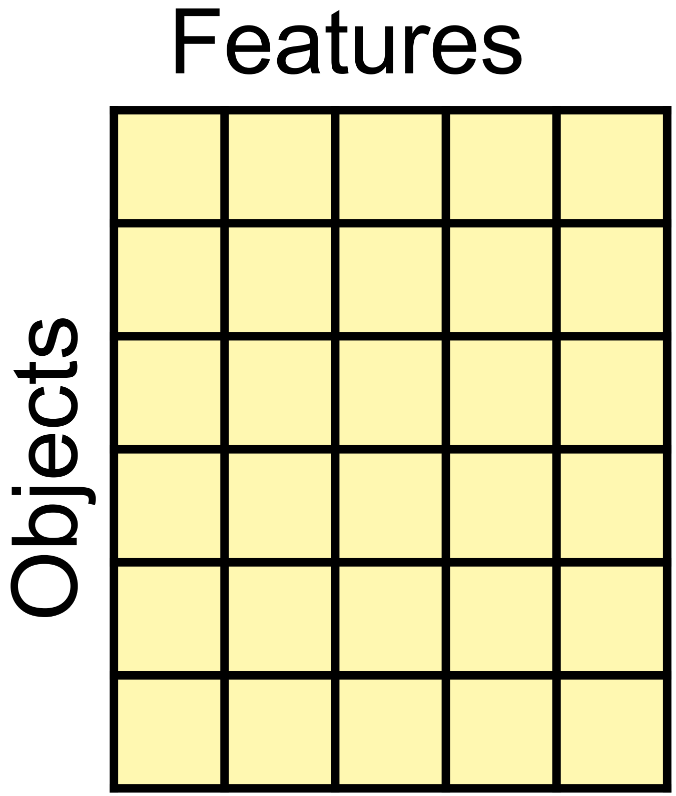


Arms race to deeply understand data

TECHNISCHE UNIVERSITÄT DARMSTADT

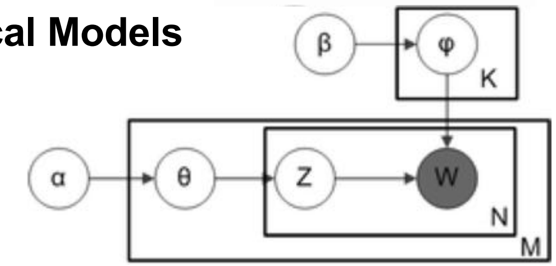


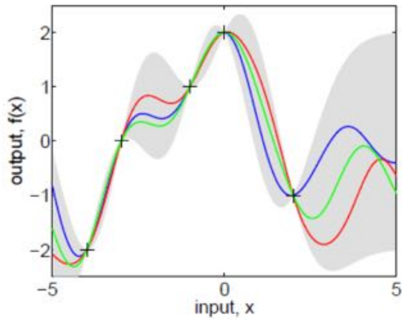
Bottom line:
Take your data spreadsheet …



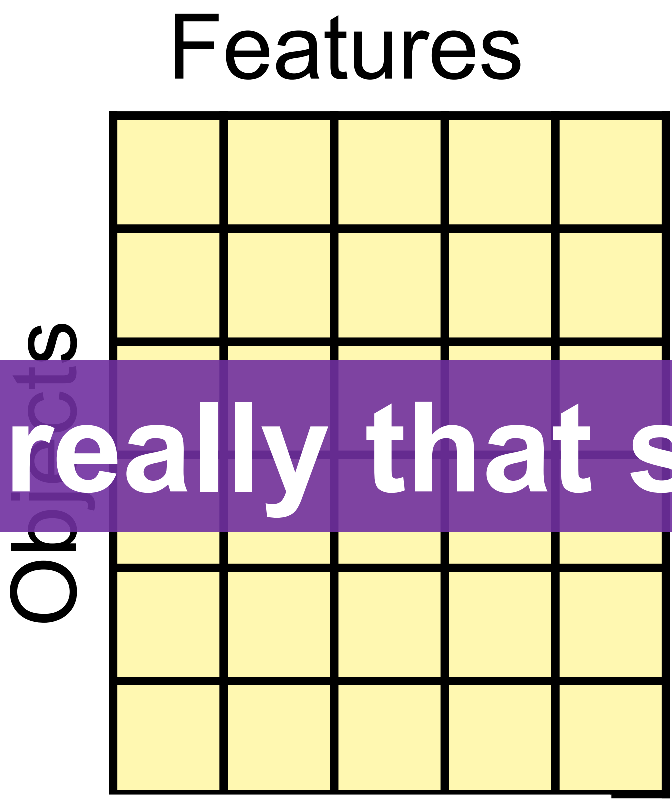
… and apply Data Science

Probabilistic Graphical Models
Arithmetic Circuits

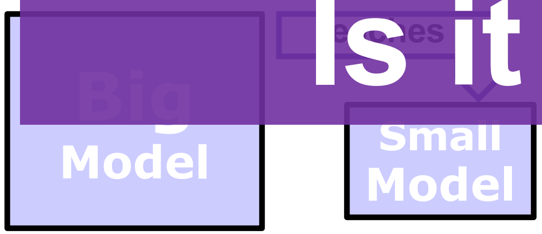




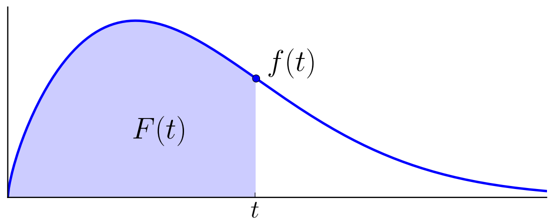
Gaussian Processes



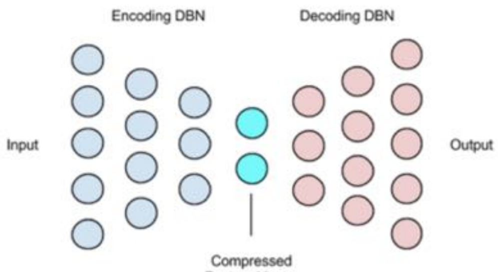
Is it really that simple?



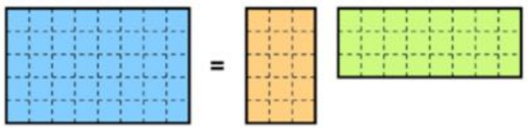
Distillation/LUPI



Diffusion Models

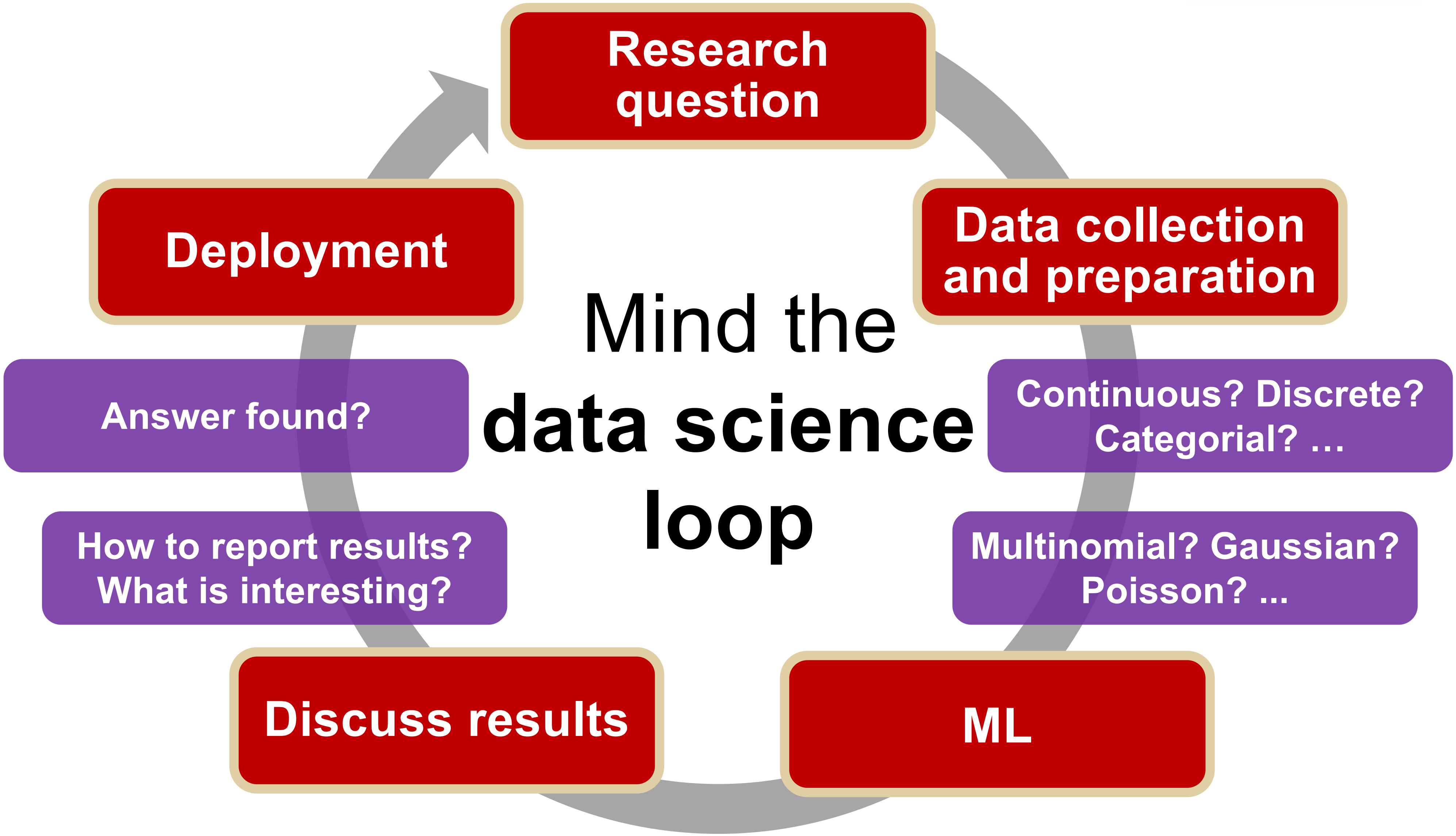


Autoencoder,
Deep Learning



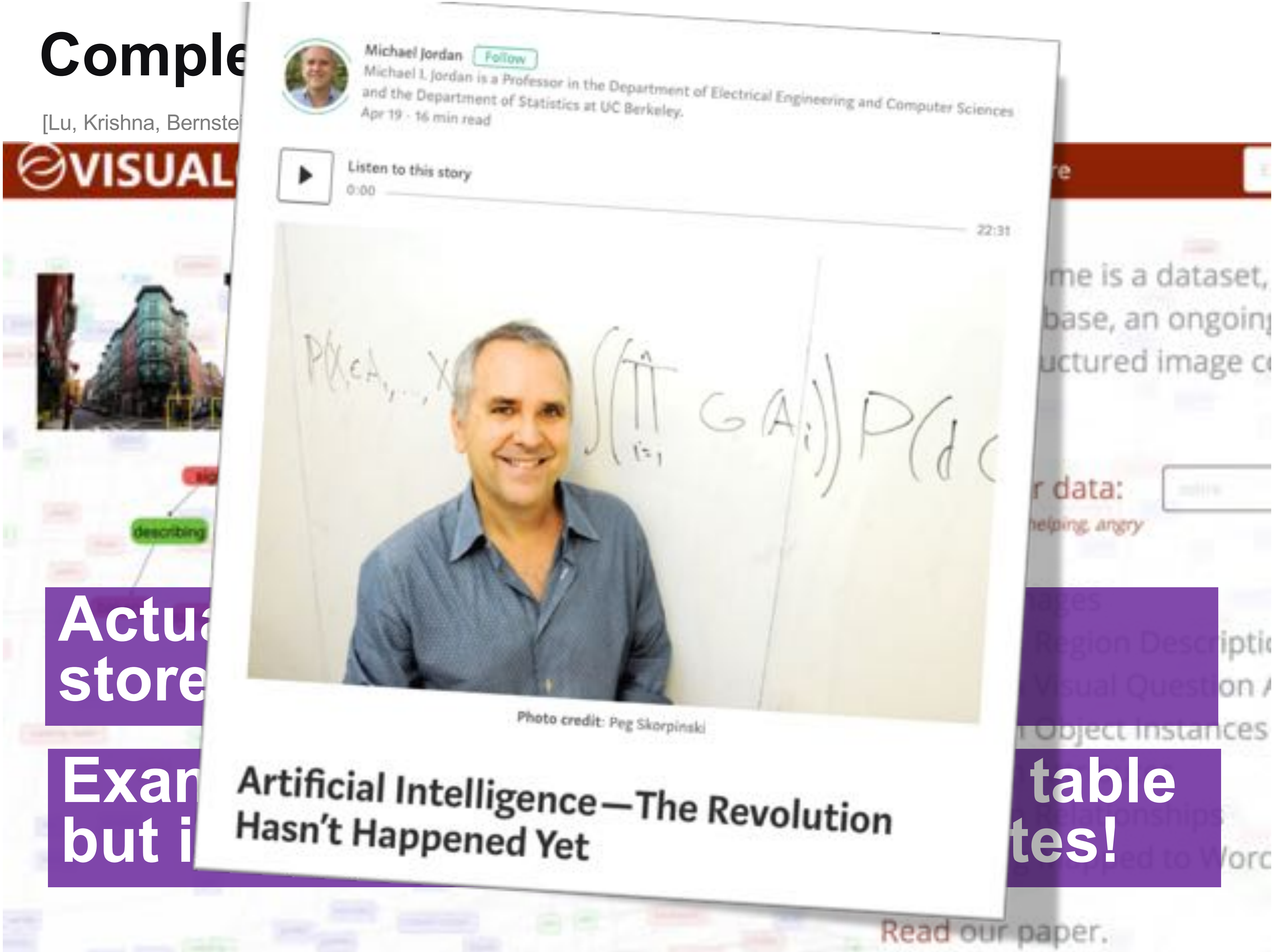
Big Data Matrix Factorization

and many more …


Research question
Data collection and preparation
Continuous? Discrete? Categorial? …
Multinomial? Gaussian? Poisson? ...
ML
Discuss results
How to report results? What is interesting?
Answer found?
Deployment
Mind the data science loop

Comple

[Lu, Krishna, Bernste



Actua
store

Exan
but i
table
tes!

We have to **democratize AI, Machine Learning, and Data Science**

We have to work on **Systems AI**, so that we know how to rapidly combine, deploy, and maintain algorithms

So yes, today is the golden era of data …

… for the best-trained, best-funded Machine Learning and Artificial Intelligence teams

Systems AI: the computational and mathematical modeling of complex AI systems.

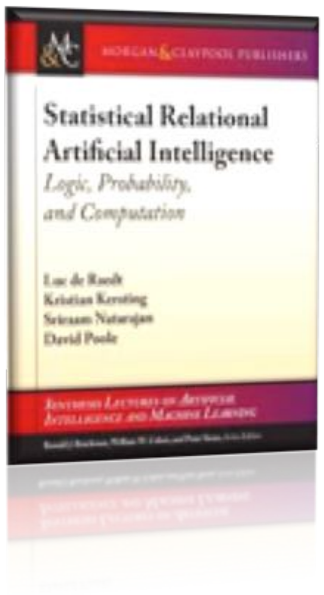
Eric Schmidt, Executive Chairman, Alphabet Inc.: Just Say "Yes”, Stanford Graduate School of Business, May 2, 2017.https://www.youtube.com/watch?v=vbb-AjiXyh0. But also see e.g. **Kordjamshidi, Roth, Kersting: “Systems AI: A Declarative Learning Based Programming Perspective.“ IJCAI-ECAI 2018.**

Part 1: For Systems AI we have to deeply understand data, knowledge and reasoning in a large number of forms

Part 2: For Systems AI we have to provide a set of tools for understanding data that require minimal expert input

Part 1: For Systems AI we have to deeply understand data, knowledge and reasoning in a large number of forms

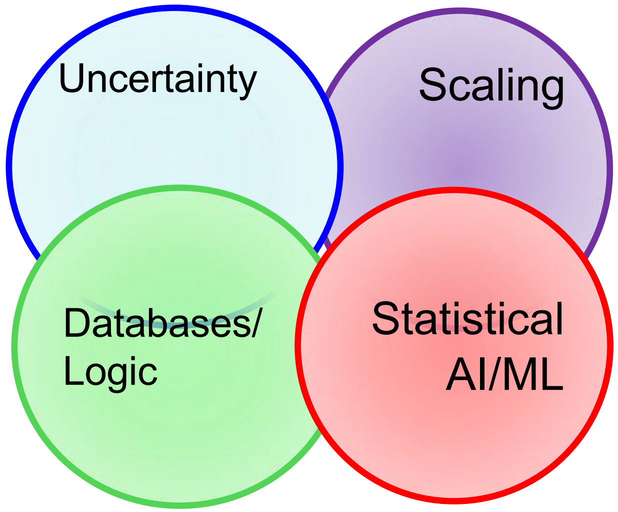
Crossover of Statistical AI/ML with data & programming abstractions



building general-purpose thinking and learning machines

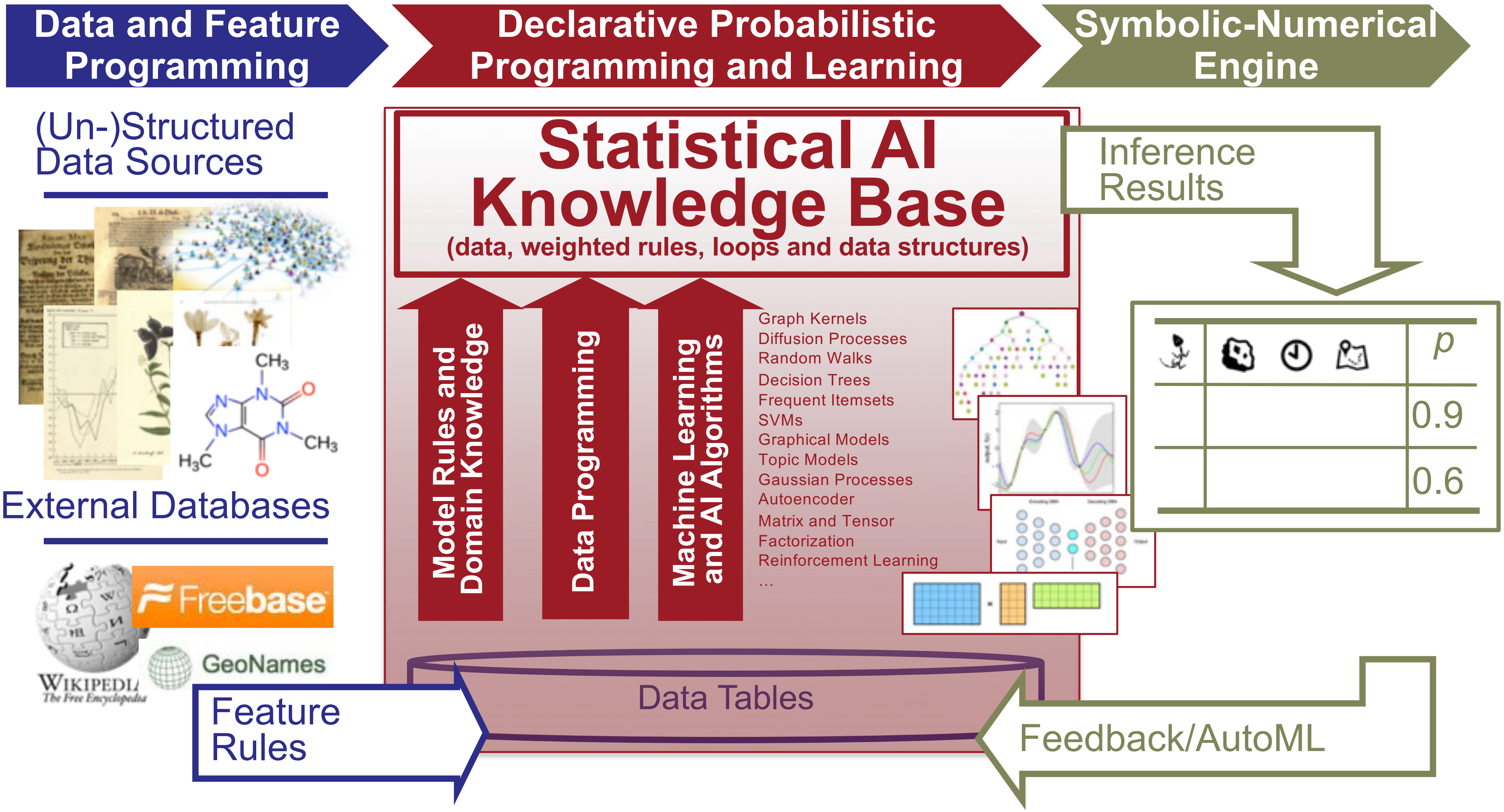
make the AI/ML expert more effective

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



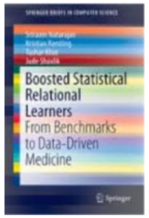
De Raedt, Kersting, Natarajan, Poole: Statistical Relational Artificial Intelligence: Logic, Probability, and Computation. Morgan and Claypool Publishers, ISBN: 9781627058414, 2016.

This establishes a novel “Deep AI”



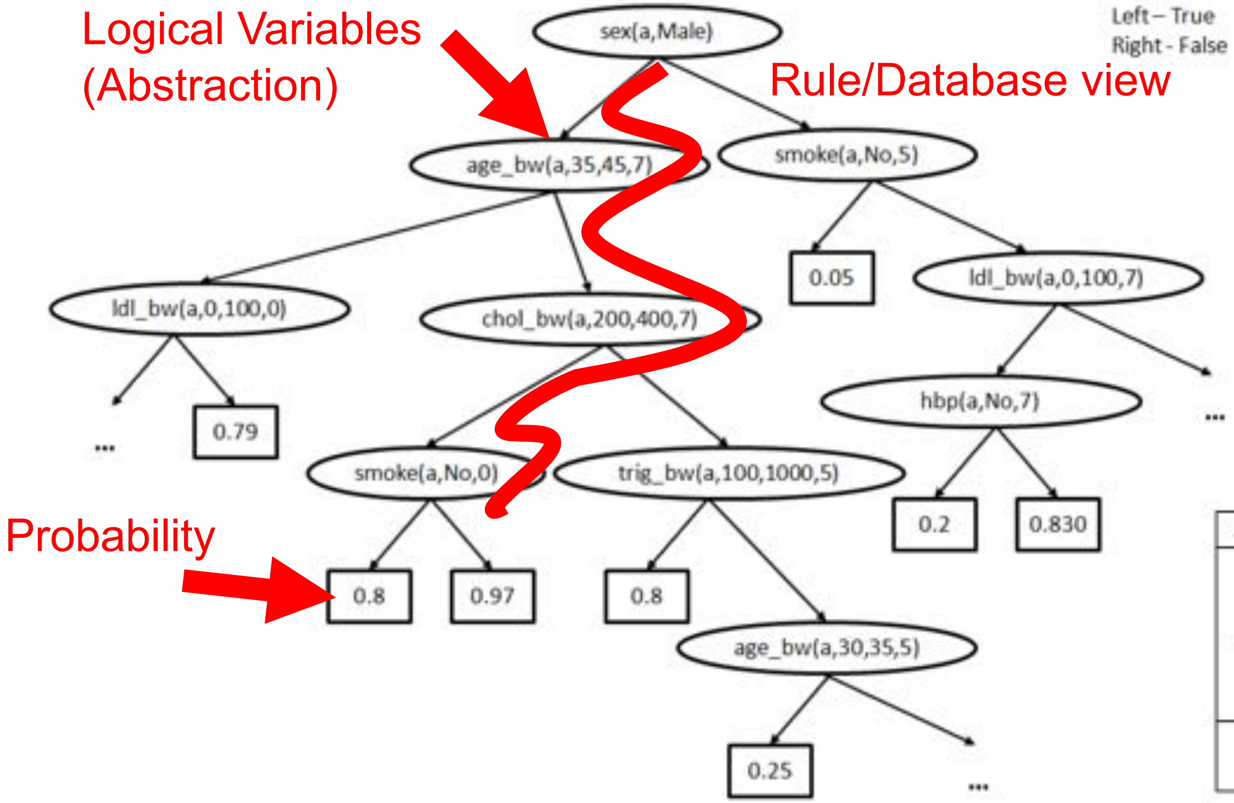
[Ré et al. IEEE Data Eng. Bull.’14; Natarajan, Picado, Khot, Kersting, Ré, Shavlik ILP’14; Natarajan, Soni, Wazalwar, Viswanathan, Kersting Solving Large Scale Learning Tasks’16, Mladenov, Heinrich, Kleinhans, Gonsior, Kersting DeLBP’16, …]

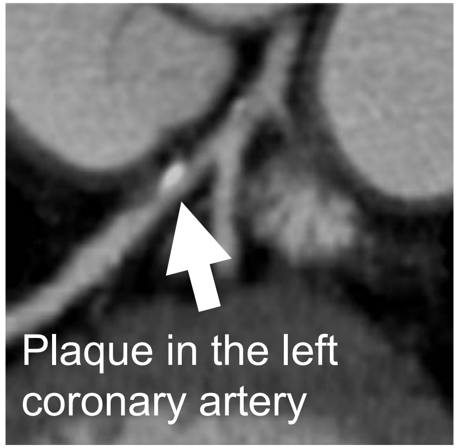
Natarajan, Khot, Kersting, Shavlik. Boosted Statistical Relational Learners. Springer Brief 2015



Mining 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
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

The higher, the better

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
Boosting	0.81	0.96	0.93	**9s**
LSM	0.73	0.54	0.62	**93 hrs**

11% | 78% | 50% | 37200x faster

[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]

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

StARLinG LAB

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

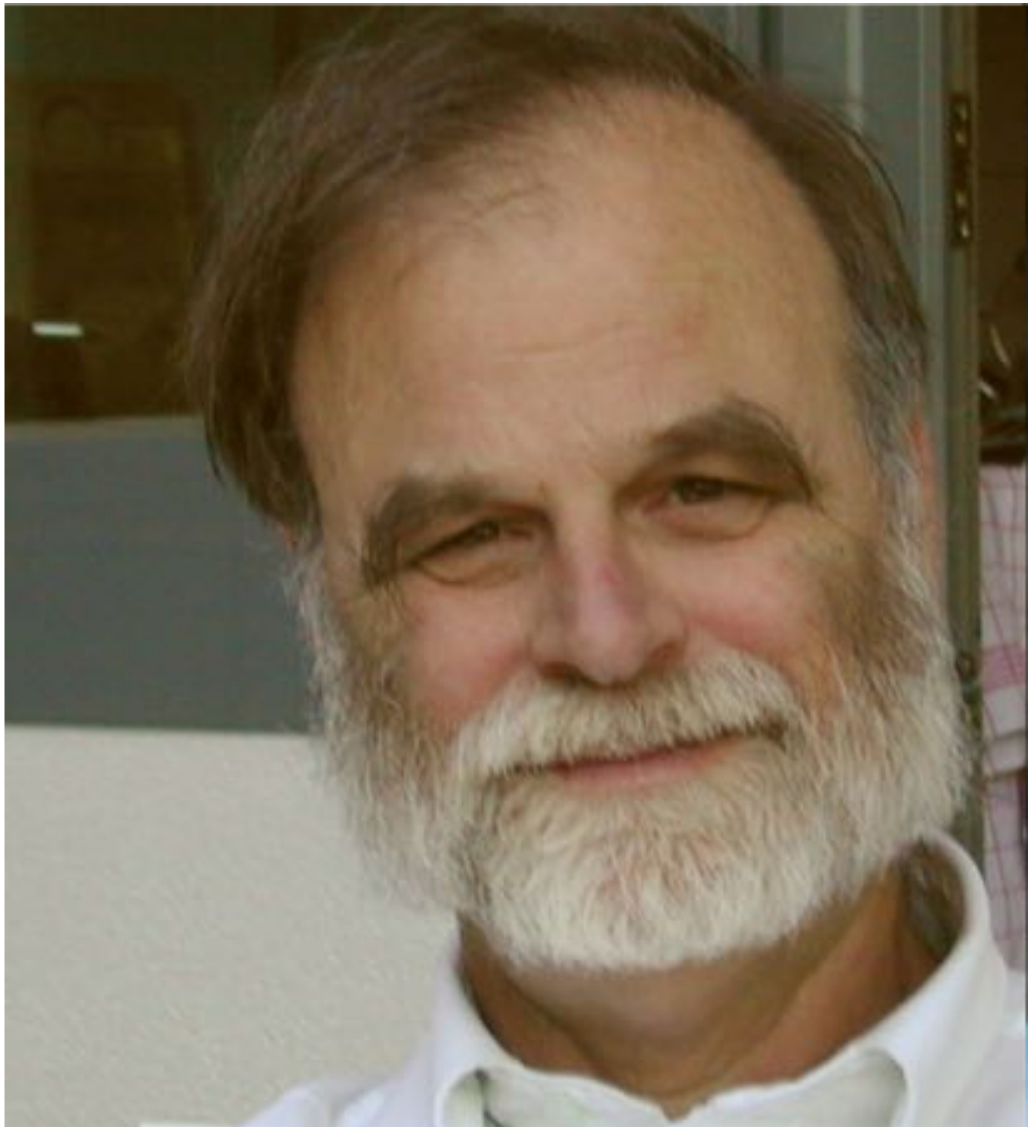
BoostSRL Wiki

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.

Sriraam 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

And connects well to database theory



Jim Gray Turing Award 1998
“Automated Programming”

Mike Stonebraker Turing Award 2014
“One size does not fit all”

… and cognitive science

"How do we humans get so much from so litt… at I
mean how do we …
g… rtle
d…


The New York Times

Opinion

A.I. Is Harder Than You Think

By Gary Marcus and Ernest Davis

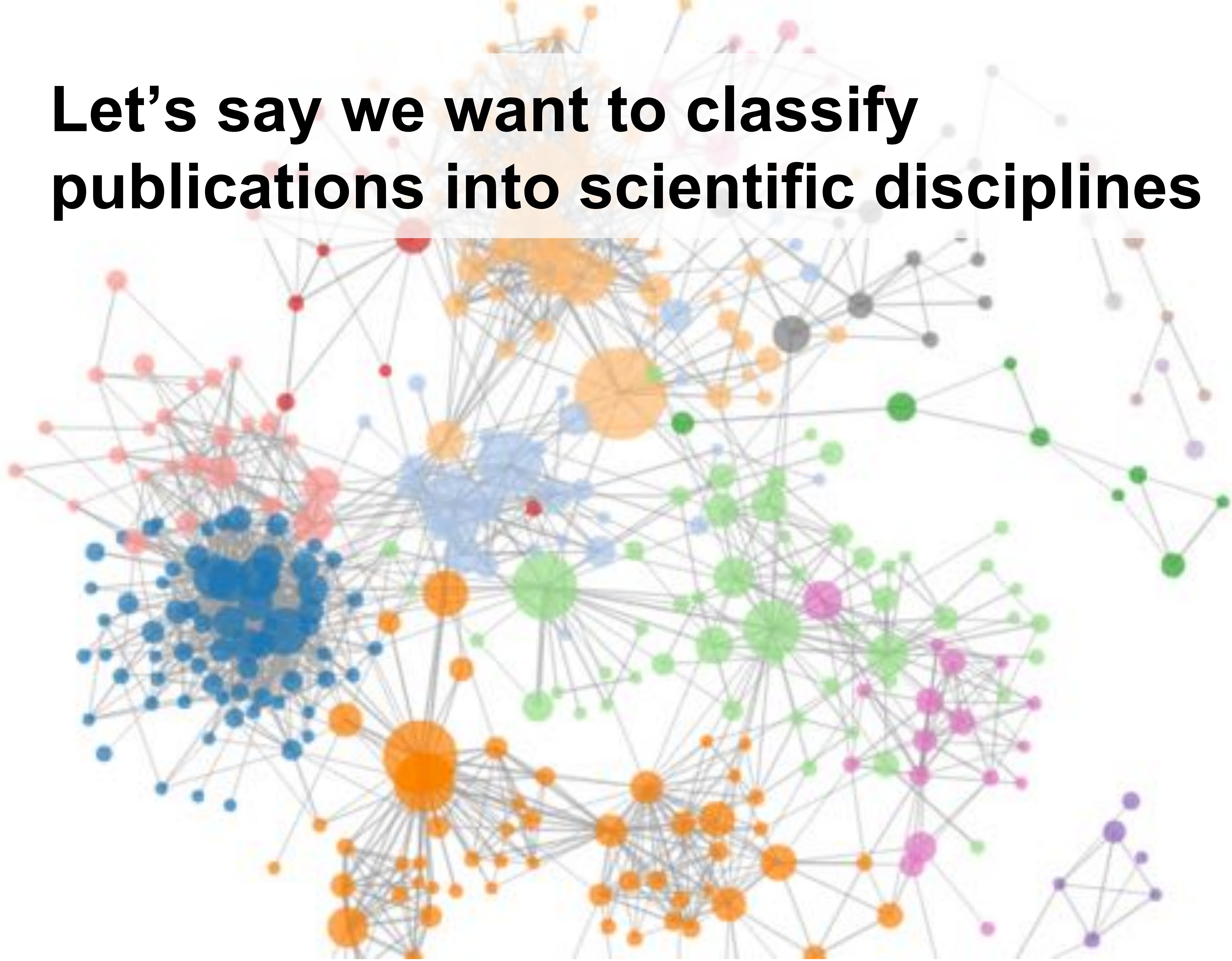
Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science.

May 18, 2018

Bayesian Program Learning”

Lake, Salakhutdinov, Tenenbaum, Science 350 (6266), 1332-1338, 2015
Tenenbaum, Kemp, Griffiths, Goodman, Science 331 (6022), 1279-1285, 2011

Let's say we want to classify publications into scientific disciplines

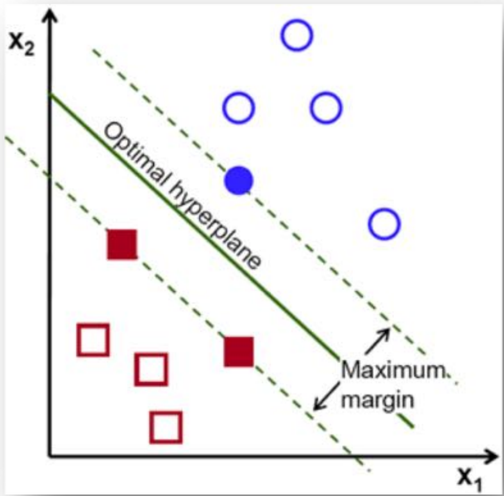


$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathcal{P}(\mathbf{w},b,\boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \begin{cases} \forall i \quad y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i)+b) \geq 1-\xi_i \\ \forall i \quad \xi_i \geq 0 \end{cases}$$

Support Vector Machines

Cortes, Vapnik MLJ 20(3):273-297, 1995



Relational Data and Program Abstractions

TECHNISCHE UNIVERSITÄT DARMSTADT

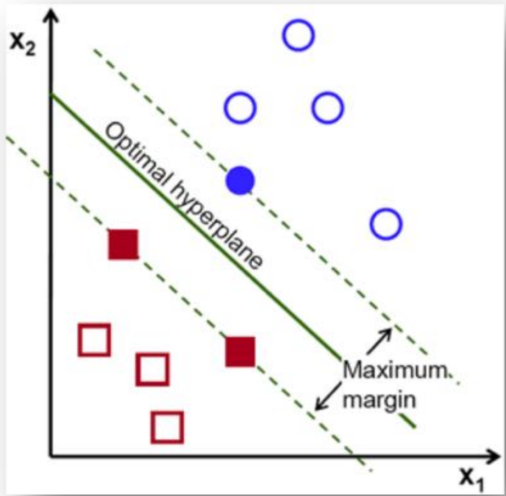
Write down SVM in „paper form.“ The machine compiles it into solver form.

```
#QUADRATIC OBJECTIVE
minimize: sum{J in feature(I,J)} weight(J)**2 + c1 * slack + c2 * coslack;

#labeled examples should be on the correct side
subject to forall {I in labeled(I)}: labeled(I)*predict(I) >= 1 - slack(I);

#slacks are positive
subject to forall {I in labeled(I)}: slack(I) >= 0;
```

Embedded within Python s.t. loops and rules can be used

reloop

RELOOP: A Toolkit for Relational Convex Optimization



Support Vector Machines
Cortes, Vapnik MLJ 20(3):273-297, 1995

But wait, publications are citing each other. OMG, I have to use graph kernels!

REALLY?

No, just add two lines of code!

TECHNISCHE UNIVERSITÄT DARMSTADT

Write down SVM in „paper form.“ The machine compiles it into solver form.

```
#QUADRATIC OBJECTIVE
minimize: sum{J in feature(I,J)} weight(J)**2 + c1 * slack + c2 * coslack;

#labeled examples should be on the correct side
subject to forall {I in labeled(I)}: labeled(I)*predict(I) >= 1 - slack(I);

#slacks are positive
subject to forall {I in labeled(I)}: slack(I) >= 0;

#TRANSDUCTIVE PART
#cited instances should have the same labels.
subject to forall {I1, I2 in linked(I1, I2)}: labeled(I1) * predict(I2) >= 1 - slack(I1, I2)
subject to forall {I1, I2 in linked(I1, I2)}: coslack(I1, I2) >= 0; #coslacks are positive
```

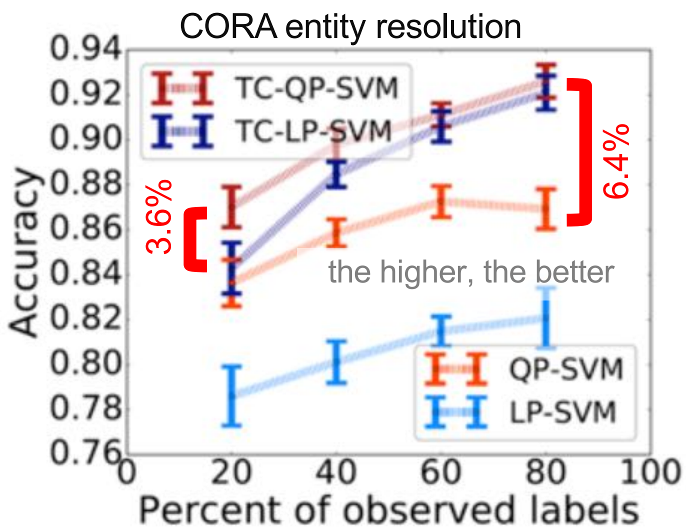
Citing papers share topics

No kernel, the structure is expressed within the constraints!

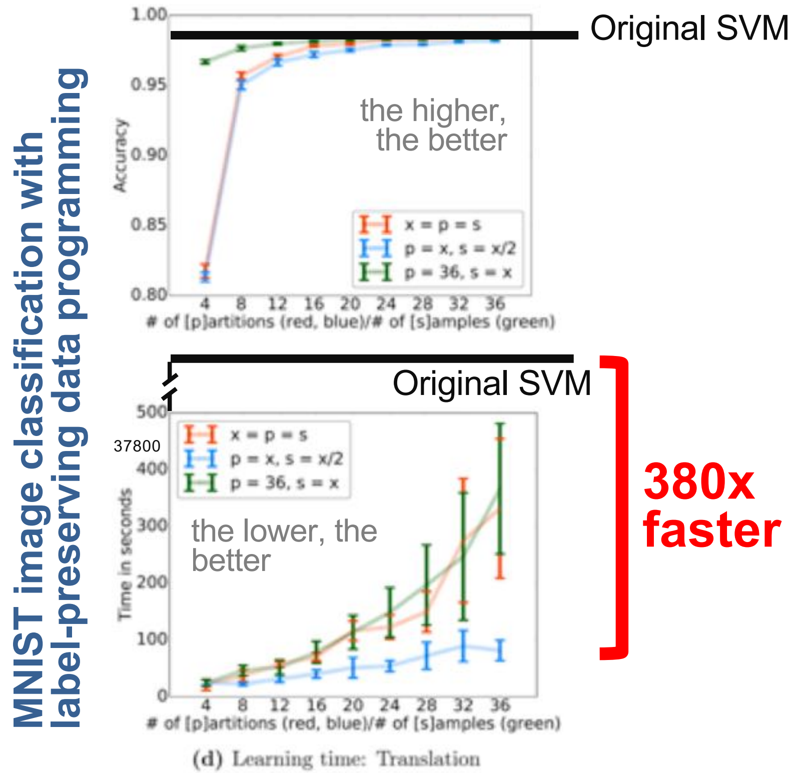
Grohe, Kersting, Mladenov, Selman ESA´14, Kersting, Mladenov, Tokmatov AIJ´17, Mladenov, Kleinhans, Kersting AAAI´17

Faster than traditional ML!

TECHNISCHE UNIVERSITÄT DARMSTADT



On par with state-of-the-art (but specialized) models by just few lines of extra code



Big Model

Run Solver

automatically compressed

Exploit computational symmetries

If exchanging two variables preserves optimality, group them together.

Small Model

Run Solver

reloop

- Overview
- Source
- Commits
- Branches
- Pull requests
- Pipelines
- Issues
- Downloads

Reloop

1. Prequisites as in requirements.txt

- Reloop requires Python 2.7+
- Scipy v0.15+
- Numpy v1.9.1+
- Cython v0.21.1+
- Cvxopt v1.1.7+
- Picos v1.0.1+
- infix v1.0.0+
- Ordered-Set v1.3.1+
- pyDatalog v0.14.6
- sympy v0.7.6+
- psycopg2 v.2.6.1+
- problog v.2.1.0.5+

If pip is available all prequisites can be installed at once by running

```
$ pip install -r requirements.txt --upgrade
```

1.1 Optional Dependencies

These optional dependencies enable additional knowledge bases for usage. While Problog and SWI-Prolog both interface Prolog, psycopg2 interface a postgres database.

- Problog v2.1+
- Psycopg2 v2.6.1+
- SWI-Prolog

2. Installation

Once all the prequisites have been installed simply run

```
python setup.py build_ext --inplace
```

followed by either

```
python setup.py install
```

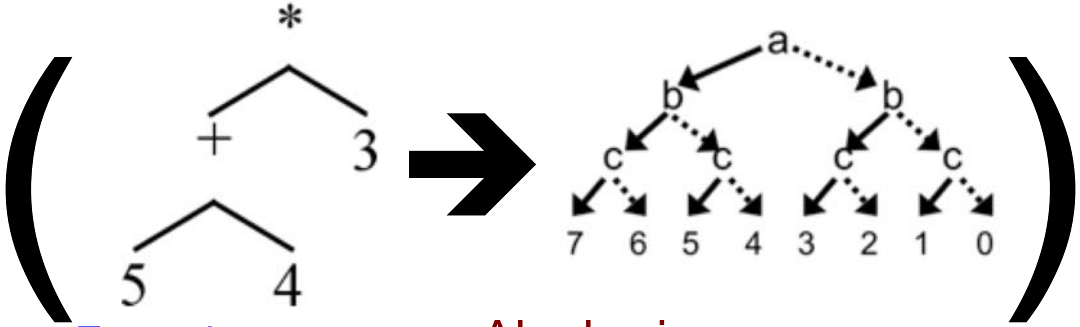
or

Embedded within Python

RELOOP: A Toolkit for Relational Convex Optimization

https://bitbucket.org/reloopdev/reloop

New field: Symbolic-numerical AI



+

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Formulae parse trees

Algebraic Decision Diagrams

Matrix Free Optimization

Problem Statistics				Symbolic IPM		Ground IPM
name	#vars	#constr	$nnz(A)$	|ADD|	time[s]	time[s]
factory	131.072	688.128	4.000.000	1819	6899	**516**
factory0	524.288	2.752.510	15.510.000	1895	**6544**	7920
factory1	2.097.150	11.000.000	59.549.700	2406	**34749**	159730
factory2	4.194.300	22.020.100	119.099.000	2504	**36248**	$\geq$ 48hrs.

>4.8x faster

Applies to QPs but here illustrated on MDPs for a factory agent which must paint two objects and connect them. The objects must be smoothed, shaped and polished and possibly drilled before painting, each of which actions require a number of tools which are possibly available. Various painting and connection methods are represented, each having an effect on the quality of the job, and each requiring tools. Rewards (required quality) range from 0 to 10 and a discounting factor of 0. 9 was used used

This "Deep AI " excites industry:

LogicBlox, Apple and Uber are investing hundreds of millions of dollars

And appears in Industrial Strength Solvers such as CPLEX and GUROBI

IBM

CPLEX

GUROBI
OPTIMIZATION

Part 2: For Systems AI we have to provide a set of tools for understanding data that require minimal expert input

Part 2: For Systems AI we have to provide a set of tools for understanding data that require minimal expert input

The Automatic Statistician

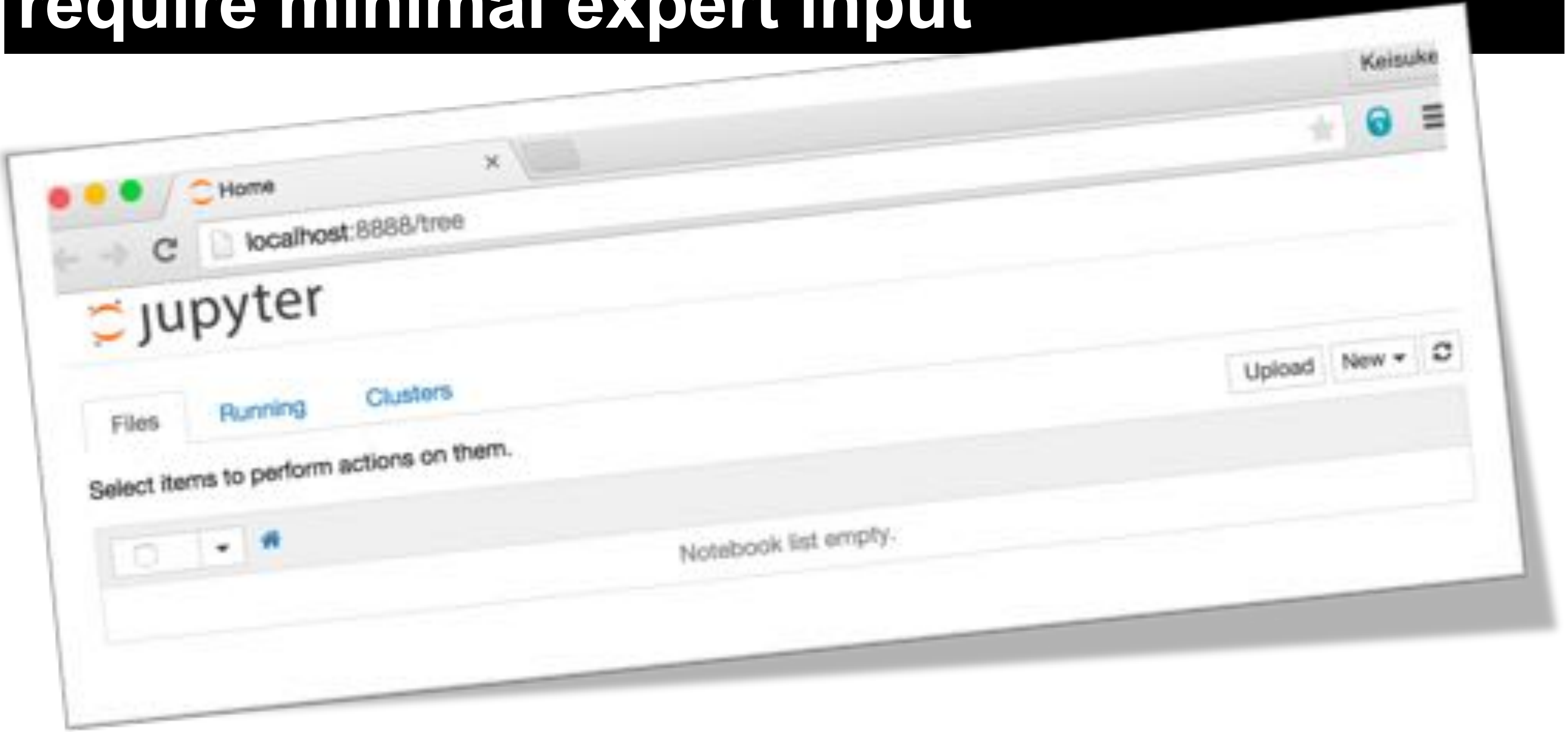
A system which explores an open-ended space of statistical models to discover a good explanation of the data, and then produces a detailed report with figures and natural-language text

No explorative data analysis yet!

Llyod, Duvenaud, Ghahramani
U. Cambridge

Grosse, Tenenbaum
MIT

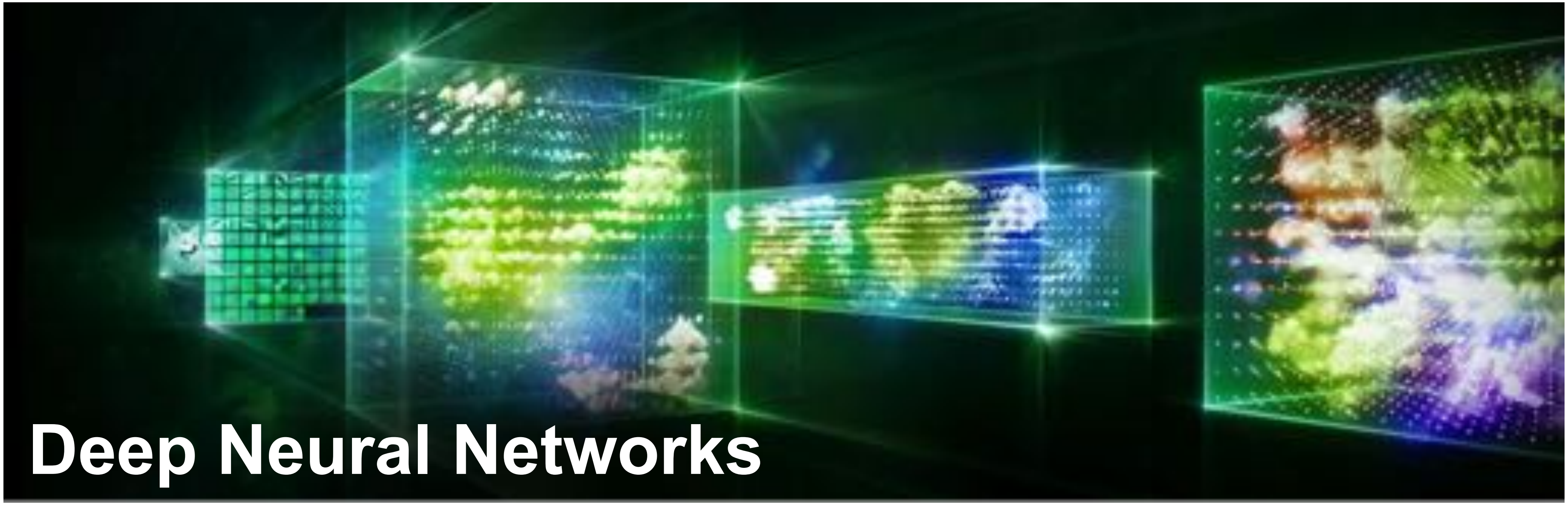
Part 2: For Systems AI we have to provide a set of tools for understanding data that require minimal expert input



Instead of starting with an empty notebook …

Part 2: For Systems AI we have to provide a set of tools for understanding data that require minimal expert input

the machine automatically compiles one for you!

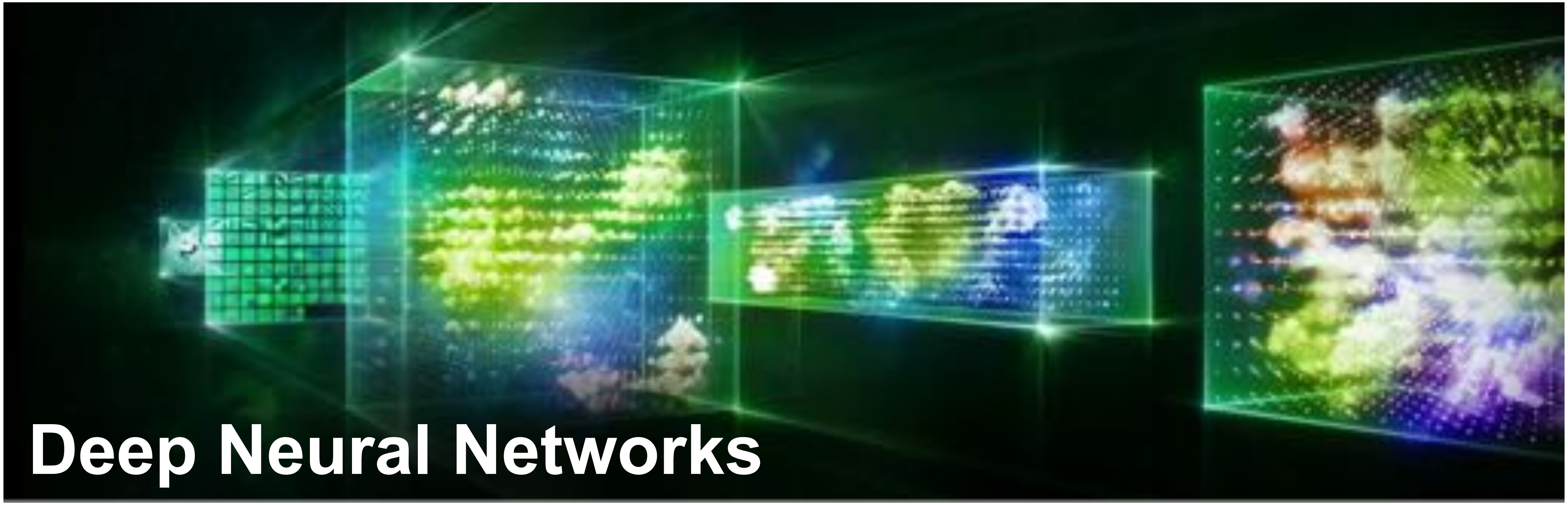


Deep Neural Networks

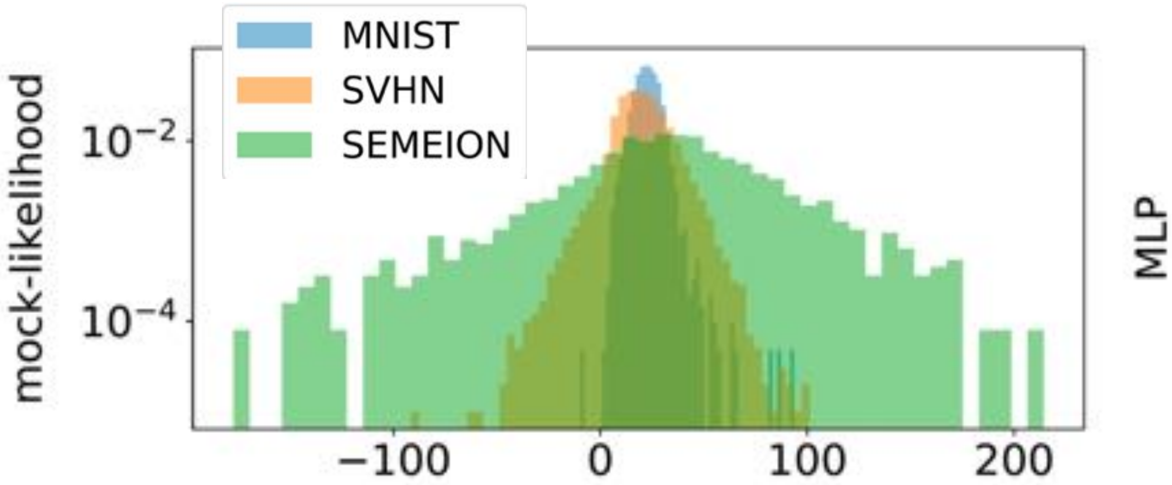
Potentially much more powerful than shallow architectures, represent computations [Bengio, 2009]

But …

- **Often no probabilistic semantics**
- **Learning requires extensive efforts**

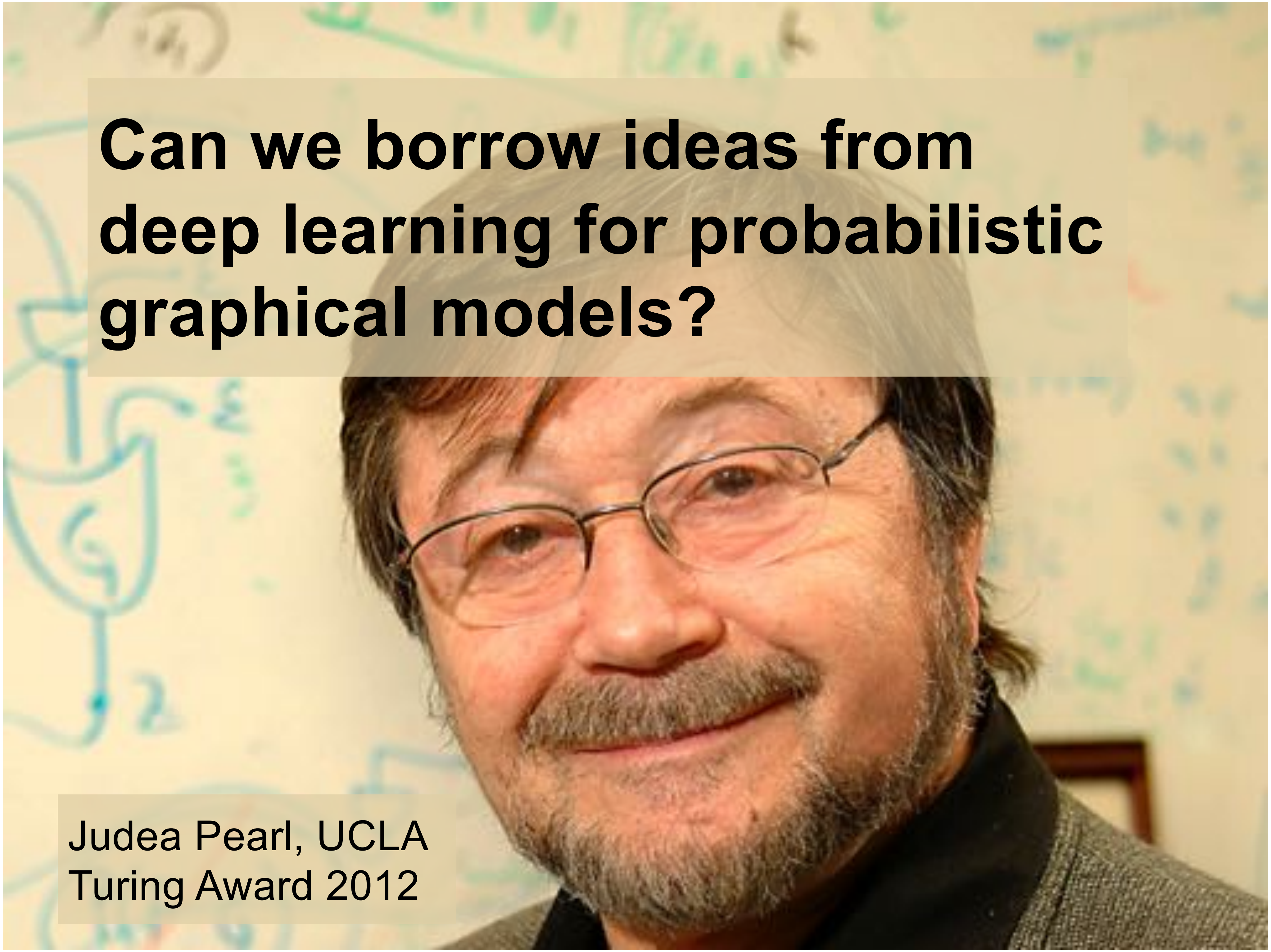


Deep Neural Networks



Deep neural networks may not be faithful probabilistic models

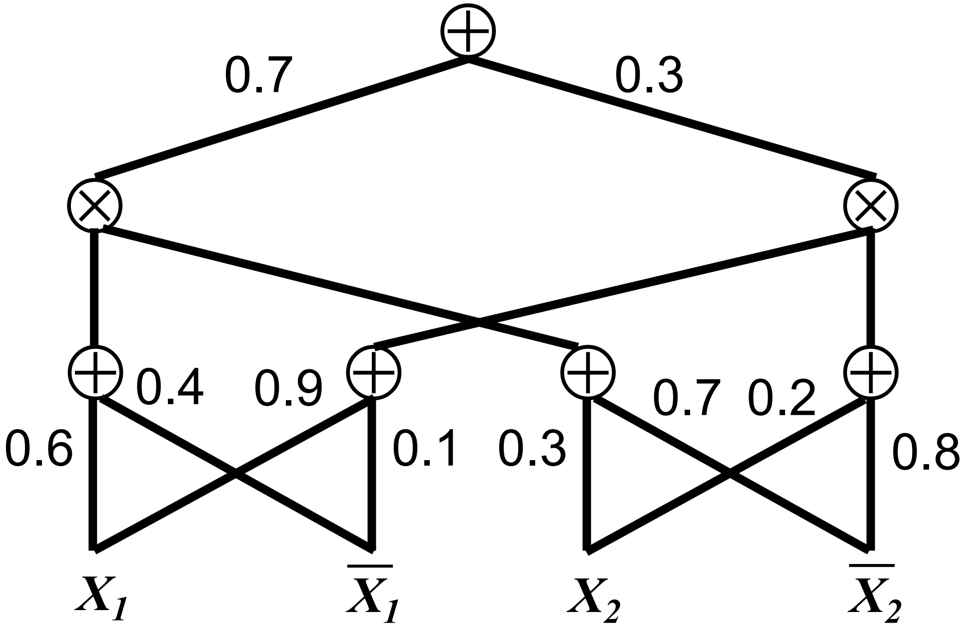
Can we borrow ideas from deep learning for probabilistic graphical models?



Judea Pearl, UCLA
Turing Award 2012

Deep Probabilistic Modelling using Sum-Product Networks

Adnan
Darwiche
UCLA



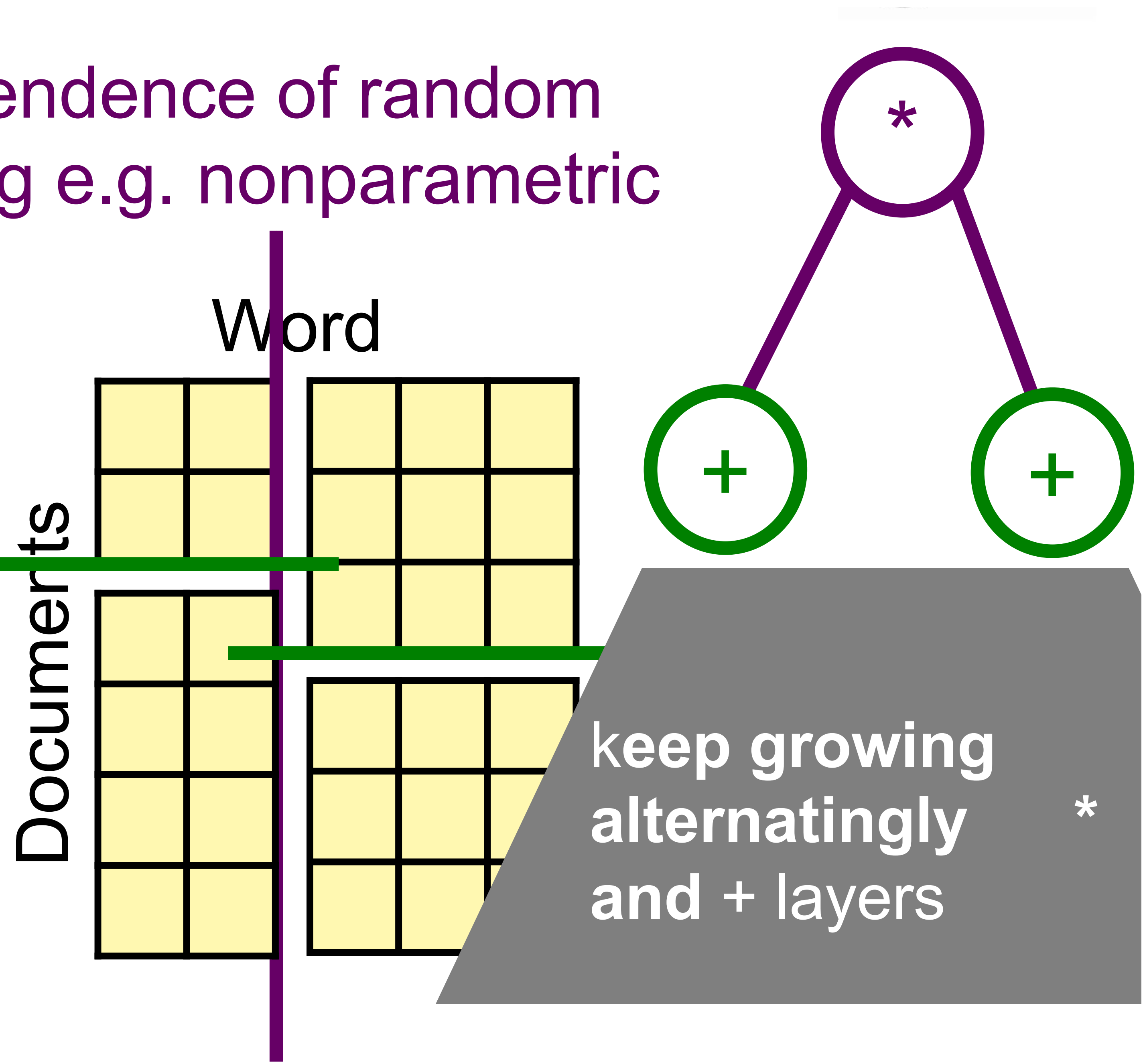
Computational graph (kind of TensorFlow graphs) that encodes how to compute probabilities

Inference is Linear in Size of Network

Greedy structure learning

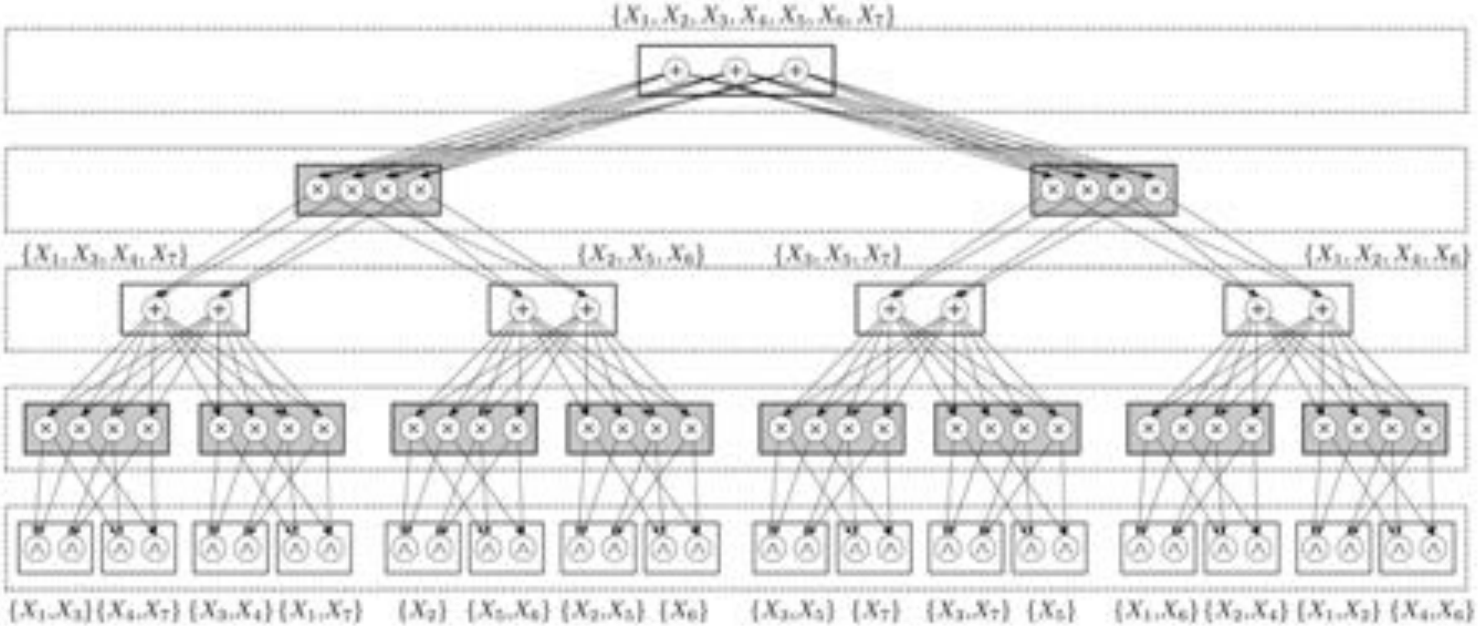
Testing independence of random variables using e.g. nonparametric tests

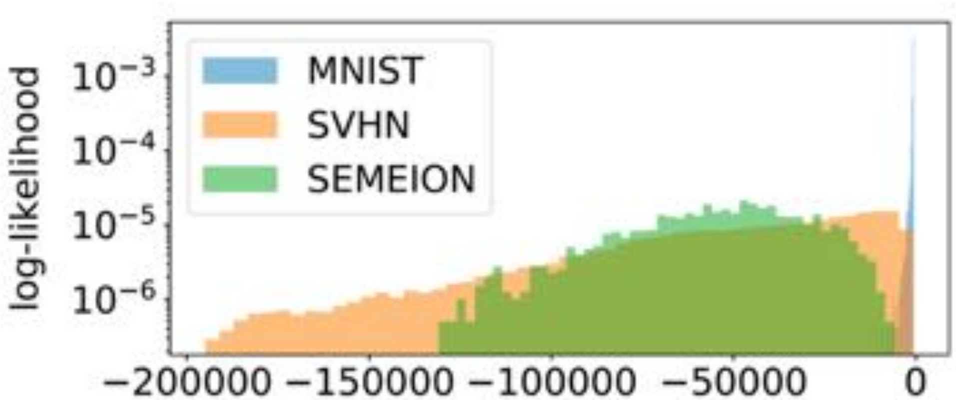
Clustering or Random splits



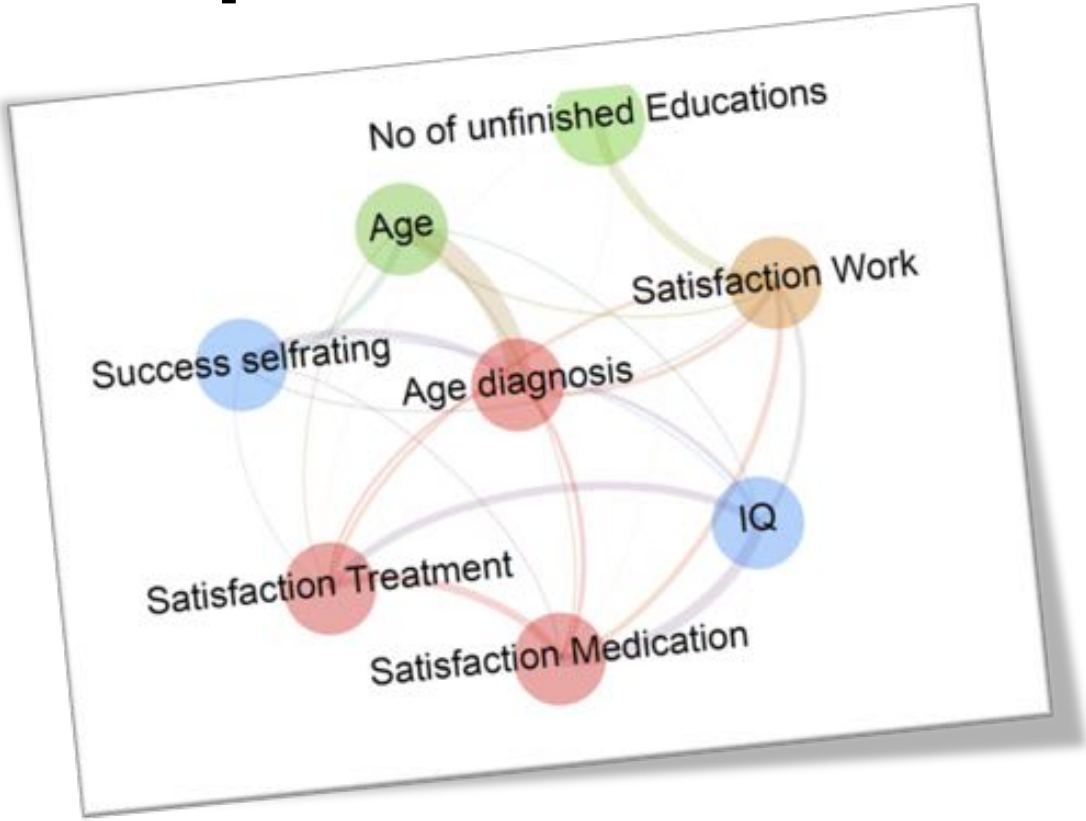
Random sum-product networks

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

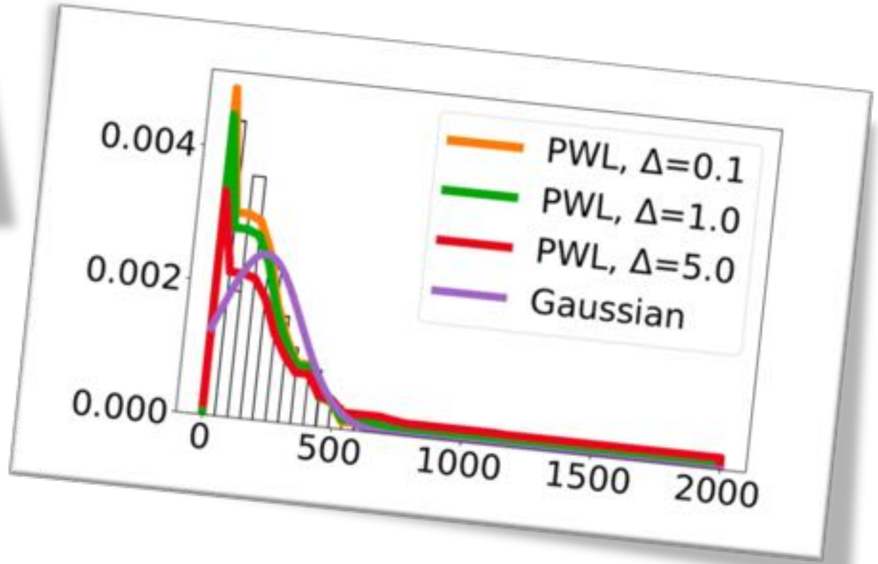




Distribution-agnostic Deep Probabilistic Learning



Use nonparametric independency tests and piece-wise linear approximations

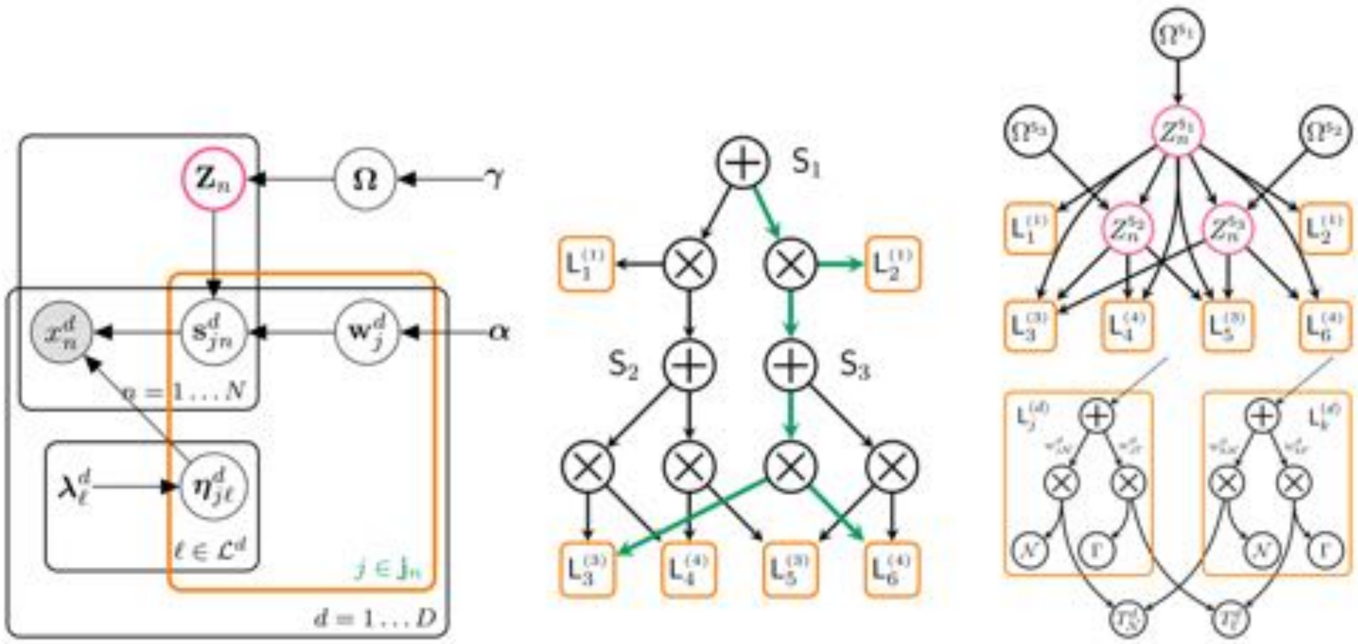


Distribution-agnostic Deep Probabilistic Learning

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? …**

Automatic Bayesian Density Analysis



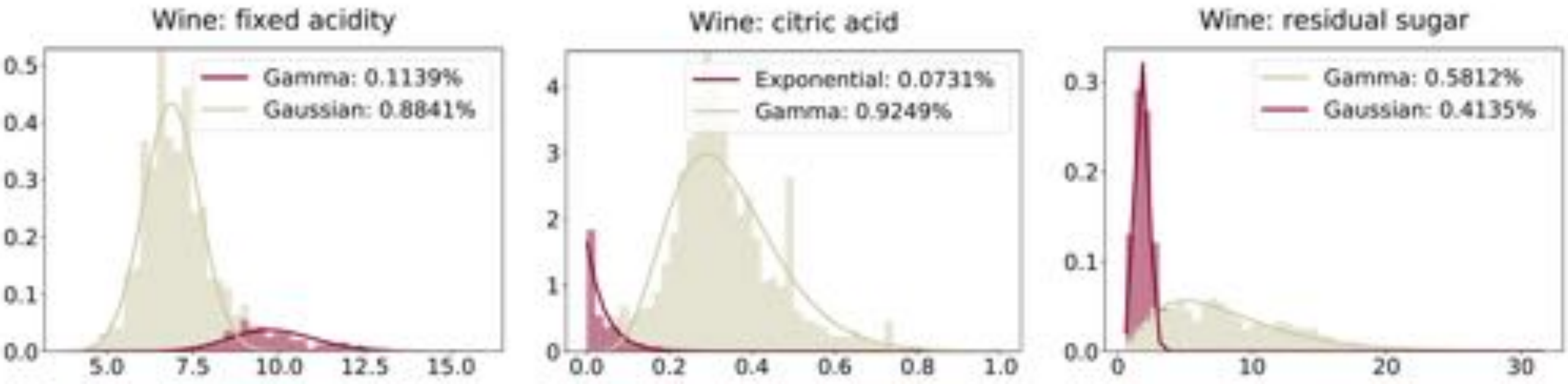
(a) Graphical model (b) SPN (c) Type-augmented SPN

Bayesian discovery of statistical types and parametric forms of variables

+

Type-agnostic deep probabilistic learning

Automatic Bayesian Density Analysis



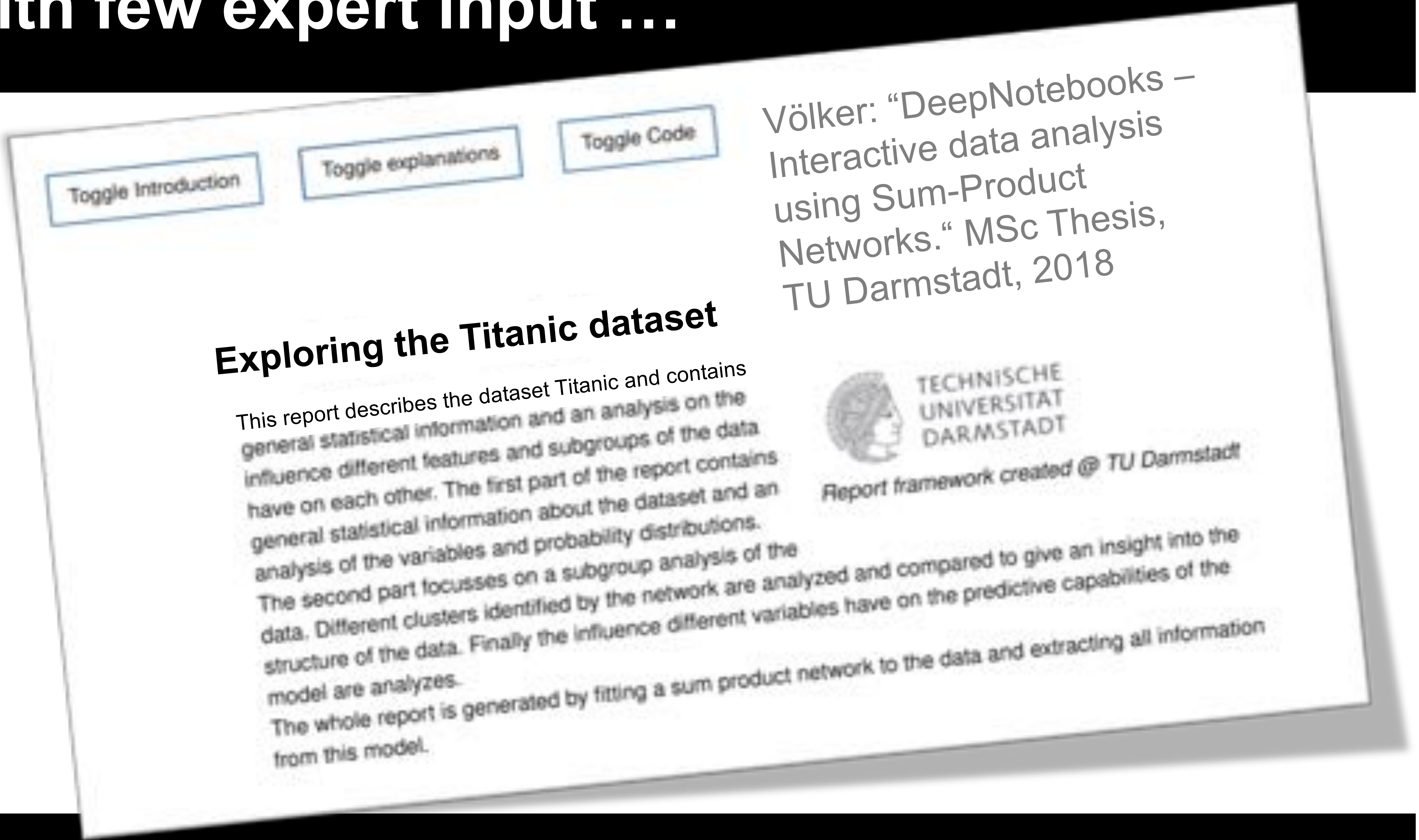
… can automatically discovers the statistical types and parametric forms of the variables

Automatic Bayesian Density Analysis

	transductive setting						*inductive setting*	
	10%			50%			70%-10%-20%	
	ISLV	**ABDA**	**MSPN**	**ISLV**	**ABDA**	**MSPN**	**ABDA**	**MSPN**
Abalone	-1.15±0.12	-0.02±0.03	**0.20**	-0.89±0.36	-0.05±0.02	**0.14**	2.22±0.02	**9.73**
Adult	-	**-0.60**±0.02	-3.46	-	**-0.69**±0.01	-5.83	**-5.91**±0.01	-44.07
Australian	-7.92±0.96	**-1.74**±0.19	-3.85	-9.37±0.69	**-1.63**±0.04	-3.76	**-16.44**±0.04	-36.14
Autism	-2.22±0.06	**-1.23**±0.02	-1.54	-2.67±0.16	**-1.24**±0.01	-1.57	**-27.93**±0.02	-39.20
Breast	-3.84±0.05	-2.78±0.07	**-2.69**	-4.29±0.17	**-2.85**±0.01	-3.06	**-25.48**±0.05	-28.01
Chess	-2.49±0.04	**-1.87**±0.01	-3.94	-2.58±0.04	**-1.87**±0.01	-3.92	**-12.30**±0.00	-13.01
Crx	-12.17±1.41	**-1.19**±0.12	-3.28	-11.96±1.01	**-1.20**±0.04	-3.51	**-12.82**±0.07	-36.26
Dermatology	-2.44±0.23	**-0.96**±0.02	-1.00	-3.57±0.32	**-0.99**±0.01	-1.01	**-24.98**±0.19	-27.71
Diabetes	-10.53±1.51	**-2.21**±0.09	-3.88	-12.52±0.52	**-2.37**±0.09	-4.01	**-17.48**±0.05	-31.22
German	-3.49±0.21	**-1.54**±0.01	-1.58	-4.06±0.28	**-1.55**±0.01	-1.60	**-25.83**±0.05	-26.05
Student	-2.83±0.27	**-1.56**±0.03	-1.57	-3.80±0.29	**-1.57**±0.01	-1.58	**-28.73**±0.10	-30.18
Wine	-1.19±0.02	-0.90±0.02	**-0.13**	-1.34±0.01	-0.92±0.01	**-0.41**	-10.12±0.01	**-0.13**
wins	0	9	3	0	10	2	10	2

… but also models its uncertainty about the statistical types and parametric forms, which can lead to better models

The machine understands the data with few expert input …



Toggle Introduction Toggle explanations Toggle Code

Exploring the Titanic dataset

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 analyzes.
The whole report is generated by fitting a sum product network to the data and extracting all information from this model.

TECHNISCHE UNIVERSITÄT DARMSTADT

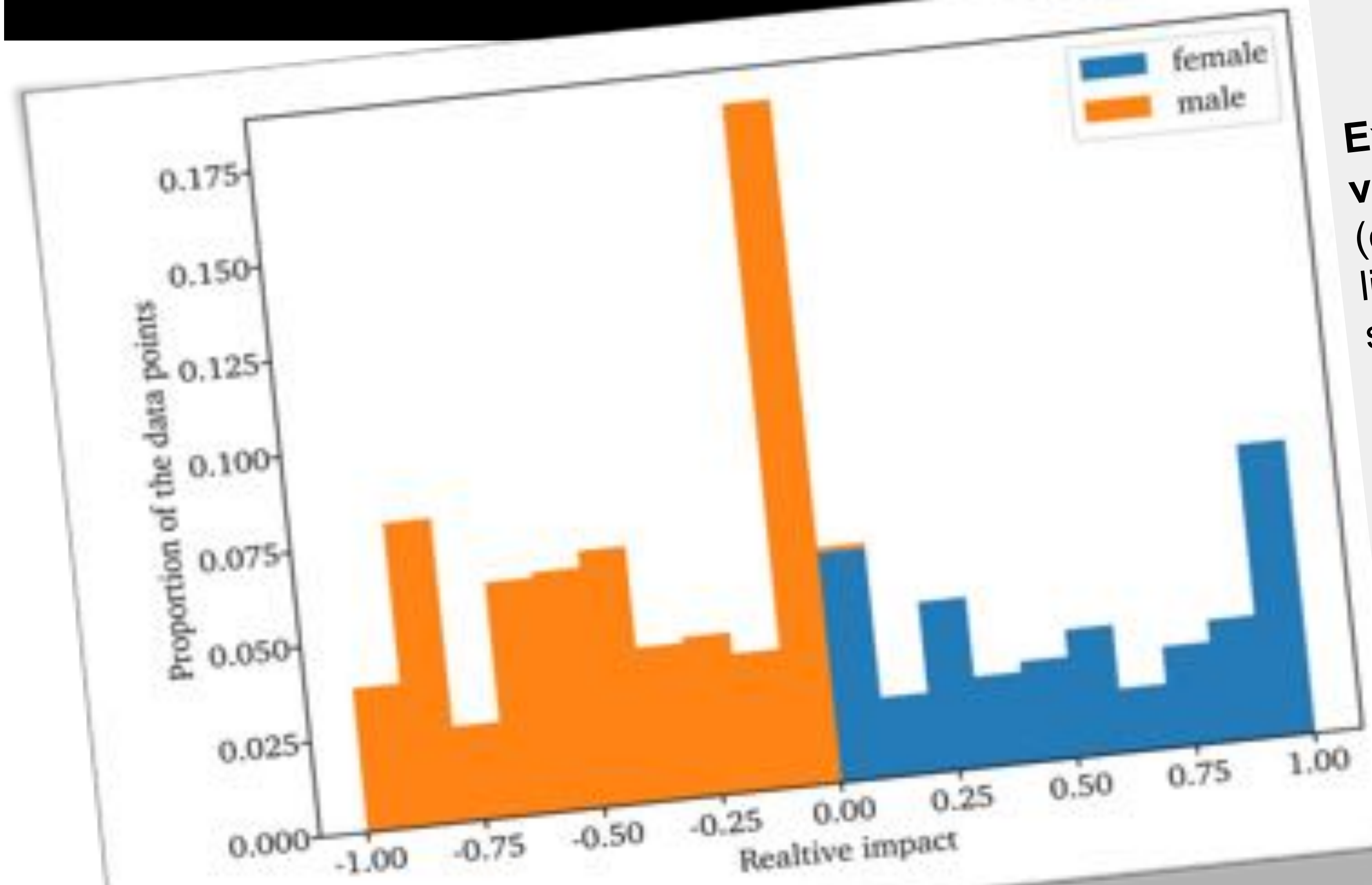
Report framework created @ TU Darmstadt

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

…and can compile data reports automatically

*[Baehrens, Schroeter, Harmeling, Kawanabe, Hansen, Müller JMLR 11:1803-1831, 2010]

The machine understands the data with no expert input …



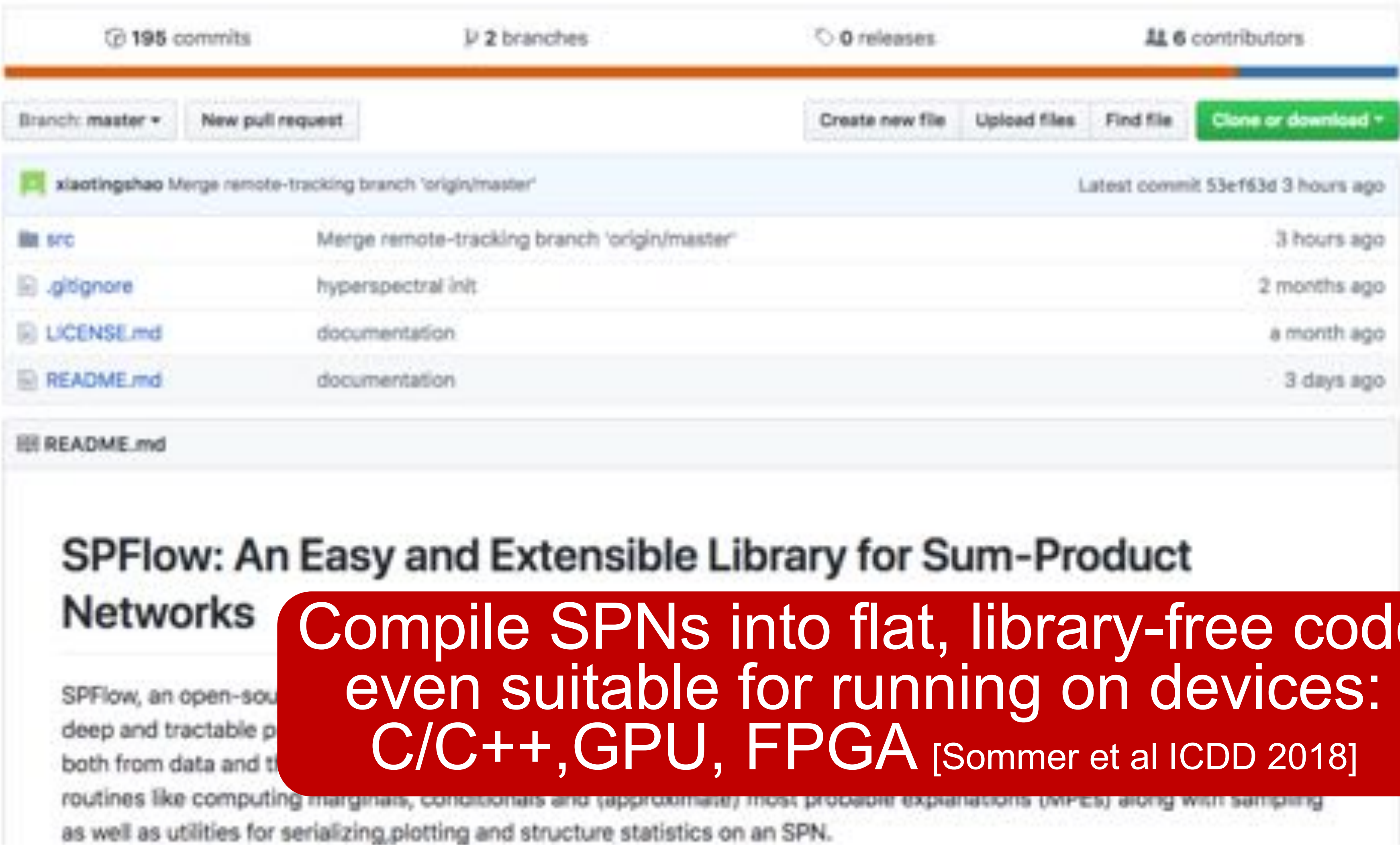
Explanation vector* (computable in linear time in the sizre of the SPN) **showing the impact of "gender" on the chances of survival for the Titanic dataset**

…and can compile data reports automatically

SPFlow: An Easy and Extensible Library for Sum-Product Networks

[Molina, Vergari, Stelzner, Peharz, Di Mauro, Kersting 2018]

https://github.com/SPFlow/SPFlow



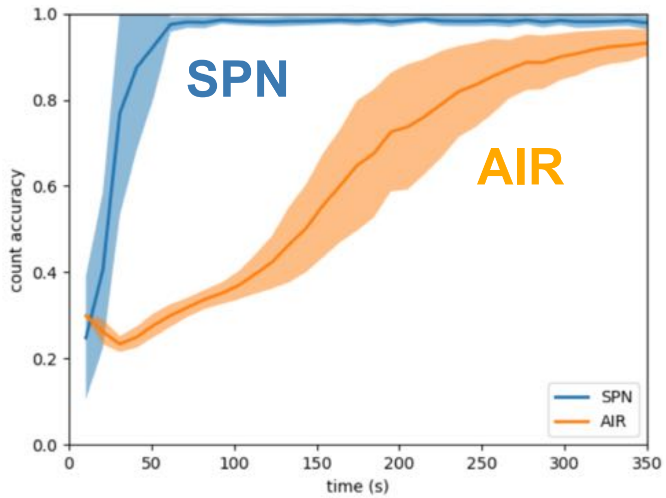
Compile SPNs into flat, library-free code even suitable for running on devices: C/C++,GPU, FPGA [Sommer et al ICDD 2018]

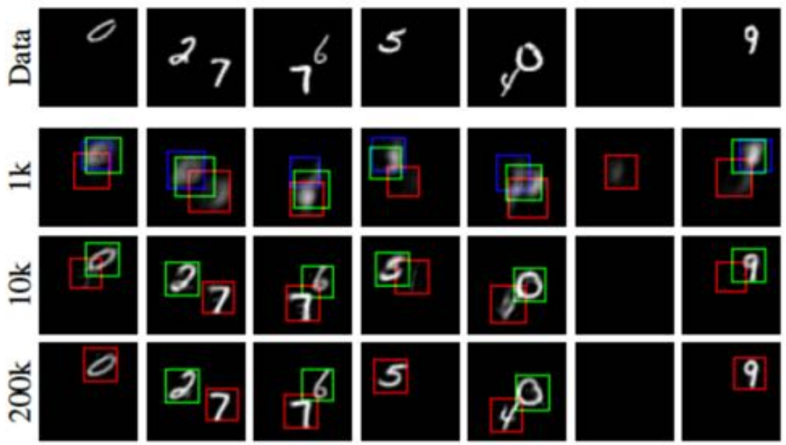
PixelSPNs

[Shao, Molina, Kersting 2018]

SPN AIR

[Stelzner, Peharz, Kersting 2018]



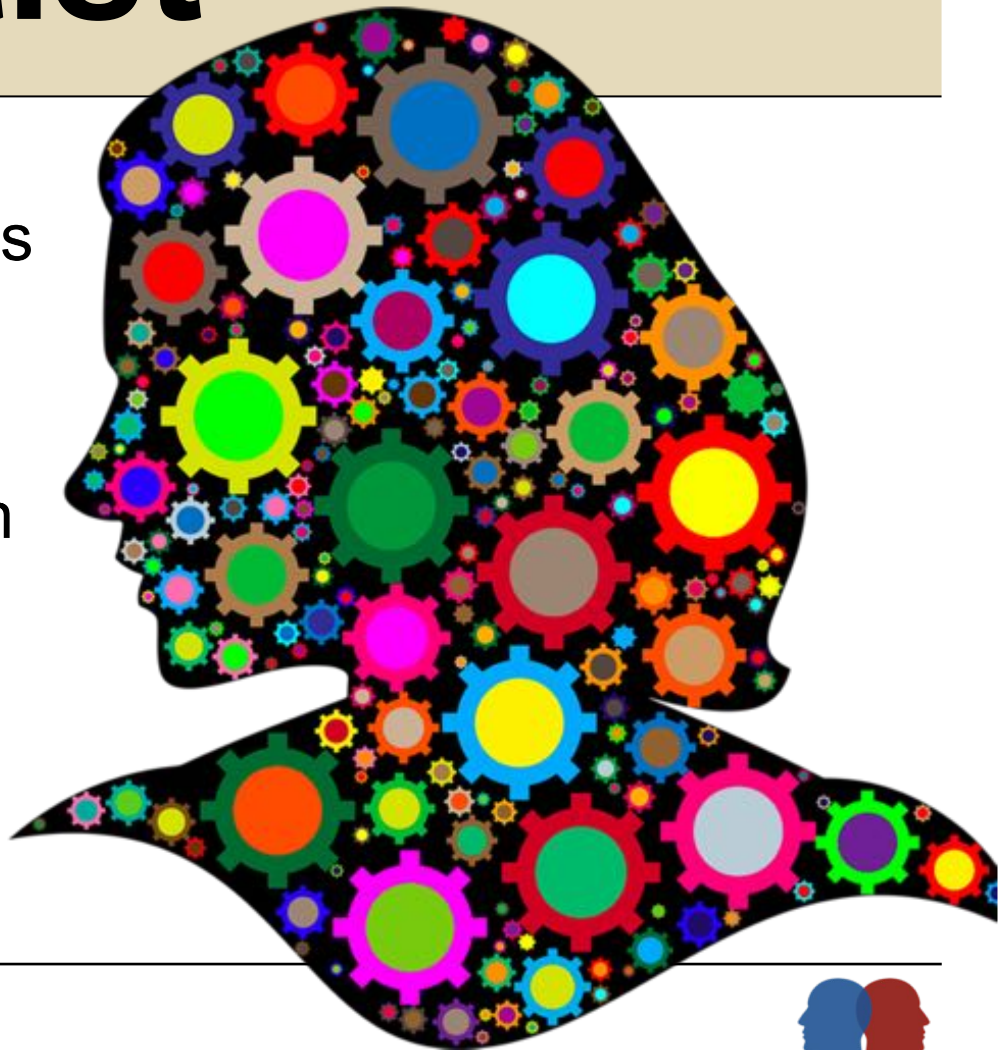


The Automatic Data Scientist

Deep probabilistic programming allows to make big steps towards making data scientists easier

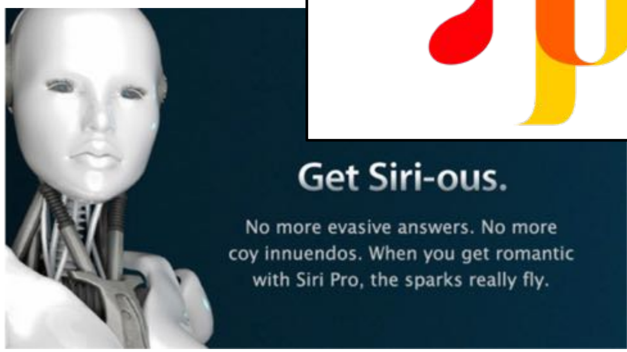
Data scientists do not have to program notebooks from scratch anymore; the machine can program major parts of them

Still a lot to be done



The Automatic Data Scientist

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



IBM CPLEX

GUROBI OPTIMIZATION

And it appears in industrial strength solvers such as CPLEX and GUROBI

Thanks for your attention