UPPSALA
UNIVERSITET

# AI for research and new results on deep regression

Thomas Schön
Uppsala University
Sweden

Chalmers AI talks
Göteborg, Sweden
June 18, 2020.

# Machine Learning – the four cornerstones

1. **Data** Typically we need lots of it.

2. **Mathematical model** A compact representation of the data that in a precise mathematical form captures the key properties.

3. **Learning algorithm** Used to compute the unknown variables from the observed data using the model.

4. **Decision/Control** Use the understanding of the current situation to steer it into a desired state.

# Key lesson from contemporary Machine Learning

**Flexible models** often give the best predictive performance.

How can we build and work with these flexible models?

1. Models that use a large (but fixed) number of parameters. (**parametric**, ex. deep learning)

   LeCun, Y., Bengio, Y., and Hinton, G. **Deep learning**, *Nature*, Vol 521, 436–444, 2015.

2. Models that use more parameters as we get access to more data. (**non-parametric**, ex. Gaussian process)

   Ghahramani, Z. **Bayesian nonparametrics and the probabilistic approach to modeling**. *Phil. Trans. R. Soc. A* 371, 2013.

   Ghahramani, Z. **Probabilistic machine learning and artificial intelligence**. *Nature* 521:452-459, 2015.

# Aim and outline

**Aim:** Motivate (using three concrete examples) the use of AI in science **and** to show a new approach for deep regression. (while at the same time keeping the discussion accessible to a general audience)

**Outline:**

1. Introduction
2. **Medicine – human-level ECG diagnosis**
3. Physics – magnetic fields
4. Biology – phylogenetics
5. Deep probabilistic regression
6. Machine Learning education

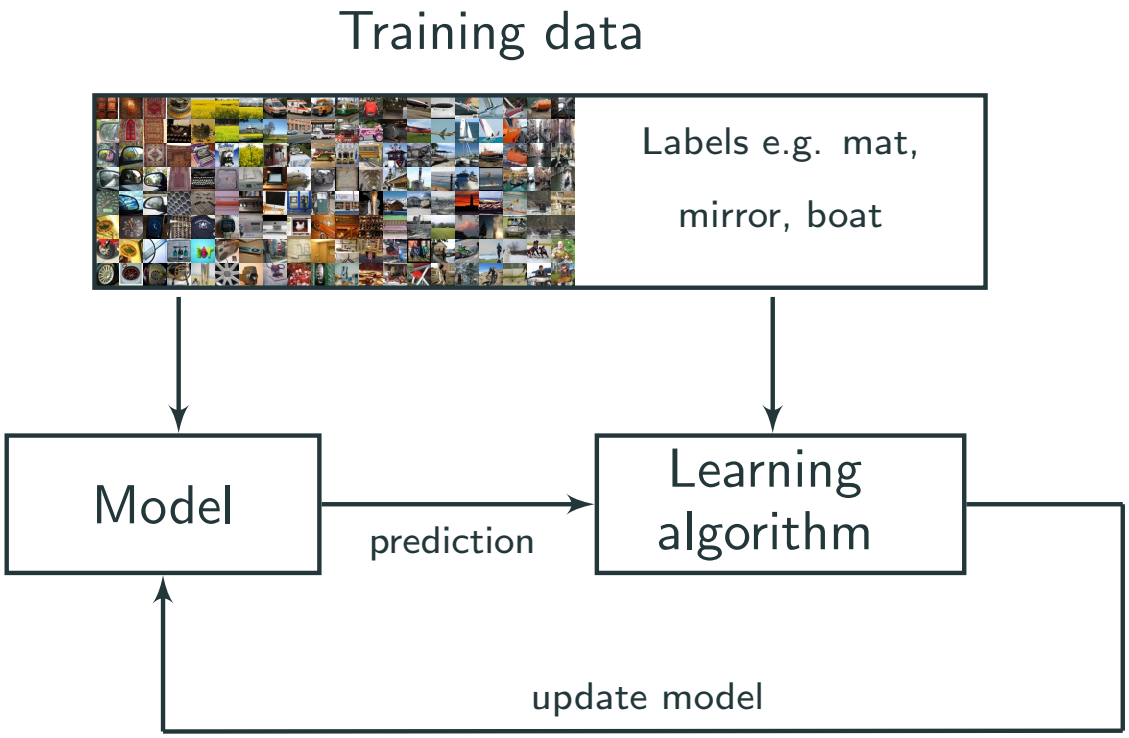
Key lesson from contemporary Machine Learning – **Flexible models often gives the best predictive performance**

# Machine learning (supervised)

Data on its own is typically useless, it is only when we can extract knowledge from the data that it becomes useful.

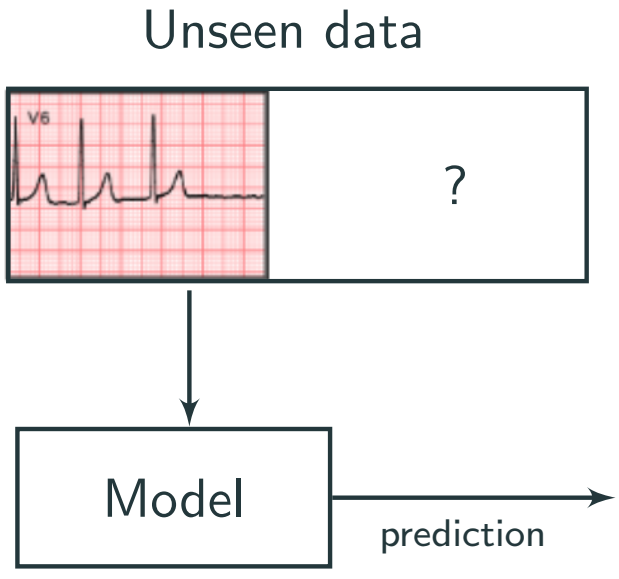
---

**Learning** a model from labelled data.



# Machine learning (supervised)

**Using** the learned model on new previously unseen data.




The model must **generalize** to new unseen data.

# Deep learning – what is it?

The mathematical model has been around for 70 years, but over the past decade there has been a **revolution**. Key reasons:

1. Very large datasets
2. Better and faster computers
3. Enormous industrial interest (e.g. Google, Facebook, MS)
4. Some methodological breakthroughs

**Underlying idea:** when representing a function, a deep, hierarchical model can be **exponentially more efficient** than a shallow model.

The functional representation has **multiple layers of abstraction**, commonly containing millions of parameters.

The parameter values are **automatically** determined based on a large amount of training data.

# Constructing a neural network for regression

A **neural network (NN)** is a hierarchical nonlinear function $y = g_\theta(x)$ from an input variable $x$ to an output variable $y$ parameterized by $\theta$.

---

**Linear regression** models the relationship between a continuous output variable $y$ and an input variable $x$,

$$y = \sum_{i=1}^{n} \theta_i x_i + \theta_0 + \varepsilon = \theta^\mathsf{T} x + \varepsilon,$$

where $\theta$ is the parameters composed by the "weights" $\theta_i$ and the offset ("bias") term $\theta_0$,

$$\theta = \begin{pmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{pmatrix}^\mathsf{T},$$

$$x = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_n \end{pmatrix}^\mathsf{T}.$$

# Generalized linear regression and NNs

We can generalize this by introducing nonlinear transformations of the predictor $\theta^{\mathsf{T}} x$,

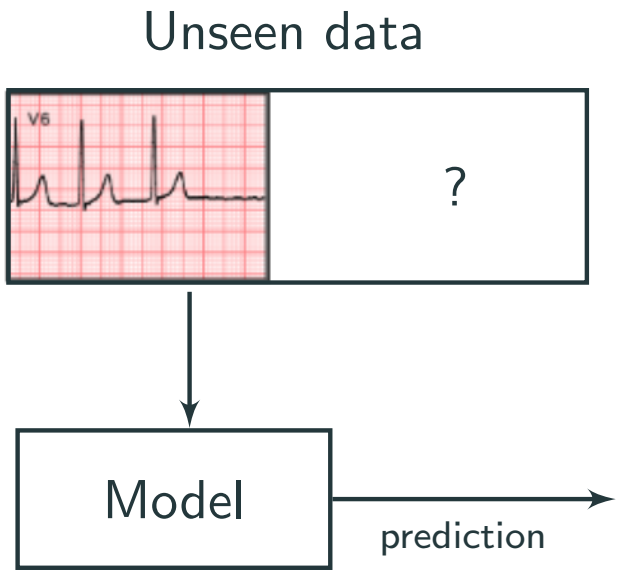
$$y = f(\theta^{\mathsf{T}} x).$$

We can think of the neural network as a **sequential construction** of several generalized linear regressions.

# Medicine – ECG diagnosis

**Aim:** Predict abnormalities based on a short-duration 12-lead electrocardiogram (ECG) recording.

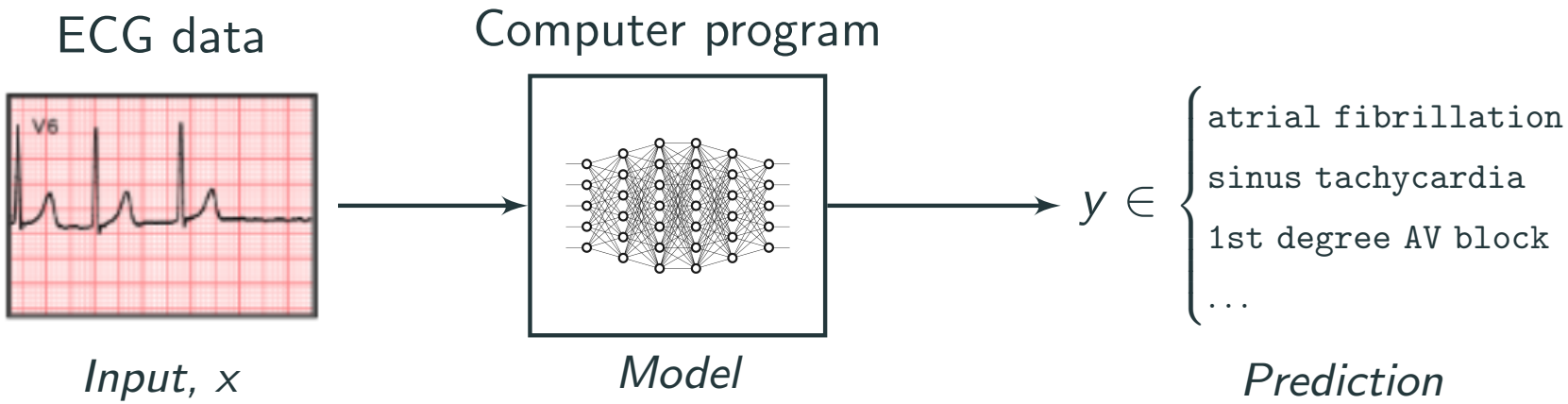
**Current situation:** The automated diagnosis that is currently available is not good enough.




**Background:** Joint work with cardiologists and ML engineers from Brazil with an urgent need for automated analysis due to the **vast distances** between the patient and a cardiologist with full expertise in ECG diagnosis.

The existing telehealth network provides the data (more than 2 300 000 ECGs), implying some clinical relevance.

# Medicine – Automatic human-level ECG diagnosis



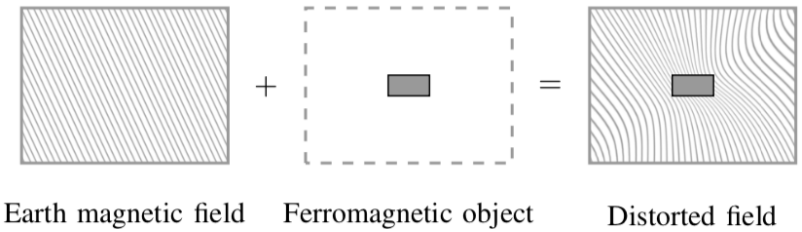

We are now reaching human level (medical doctor) performance on certain specific tasks.

**Key difference** to "classical engineering": The model is **not** derived based on our ability to mathematically explain what we see in an ECG.

Instead, a generic model is **automatically learned** based on data.

Antonio H. Ribeiro, Manoel H. Ribeiro, Gabriela M.M. Paixao, Derick M. Oliveira, Paulo R. Gomes, Jessica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., TS, Antonio Luiz P. Ribeiro. **Automatic diagnosis of the 12-lead ECG using a deep neural network**. *Nature Communications*, 11(1760), 2020.

# Physics – Ambient magnetic field map

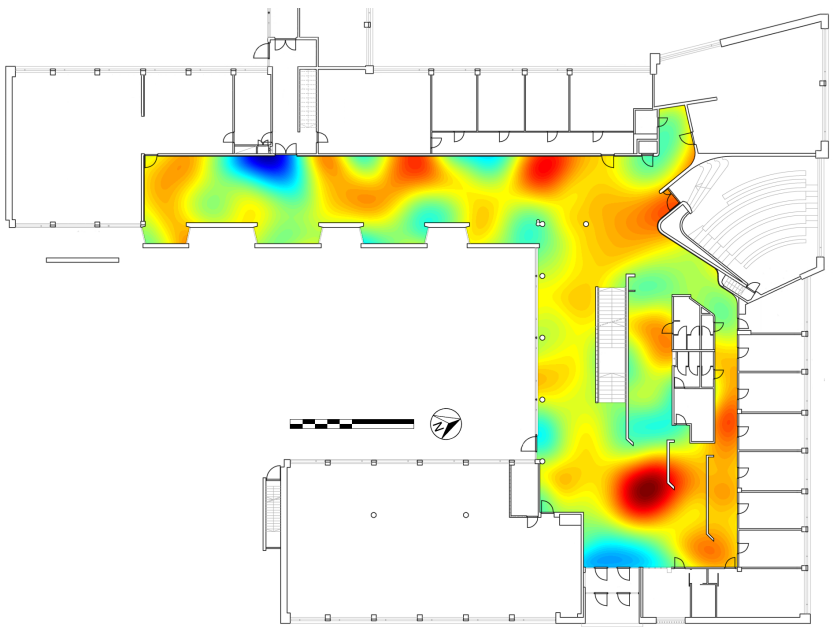



The Earth's magnetic field sets a background for the ambient magnetic field. Deviations make the field vary from point to point.

---

**Aim:** Build a map (i.e., a model) of the magnetic environment based on magnetometer measurements.

**Solution:** Customized Gaussian process that obeys Maxwell's equations.



Arno Solin, Manon Kok, Niklas Wahlström, TS and Simo Särkkä. **Modeling and interpolation of the ambient magnetic field by Gaussian processes**. *IEEE Transactions on Robotics*, 34(4):1112–1127, 2018.

Carl Jidling, Niklas Wahlström, Adrian Wills and TS. **Linearly constrained Gaussian processes**. *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, December, 2017.

# Blending prior knowledge and data

While we can do a lot with our data and flexible black-box models, we have already understood a lot about nature.

**Obvious idea: What if we could combine the two?!**

Meaning that we start from small (rigid) models describing the phenomenon we are studying and augment them with flexible models driven by data.

---

**Personal opinion:** I believe that there are (massive) gains to be maid in the simple combination of flexible data-driven models and solid widely available knowledge that we already have.

# Vision

Create flexible model building blocks **containing** the basic knowledge we have about the phenomenon we are studying.

I stress the fact that the model should be **flexible** enough to allow for new knowledge to be gained.

The data complements our existing basic knowledge and adapts it to the specific situation we are studying.

Has the potential to also allow us to learn new basic knowledge.

**Reflection:** Quite obvious really, but surprisingly little has been done...

I foresee such building blocks containing basic knowledge about physics, chemistry, psychology, biology, etc.

# Blending prior knowledge and data

**Resulting technical challenge:** How can we use known structures and domain knowledge to design priors?

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Once we have designed such a prior it will effectively **restrict the flexibility in a goal-oriented fashion**.

**Question:** What is the right blend of such priors and data?

# Starting somewhere – linear functional constraints

**Fact:** Linear functional constraints and measurements are **useful** in describing nature and **simple** to work with.

Very specific examples:

1. The magnetic field $H$ is curl-free (recall example from before)

$$\nabla \times H = 0.$$

2. Measurements are expressed as line integrals of the target function
   - X-ray computed tomography (CT)
   - Strain field reconstruction from neutron diffraction experiments

Carl Jidling, Niklas Wahlström, Adrian Wills and TS. **Linearly constrained Gaussian processes**. *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, December, 2017.

Johannes Hendriks, Carl Jidling, Adrian Wills, TS. **Linearly constrained neural networks**. *arXiv:2002.01600*, March, 2020.

# Biology – Phylogenetics via probabilistic programming

**Background/Motivation:** In statistical phylogenetics, we are interested in learning the parameters of models involving **evolutionary trees**.

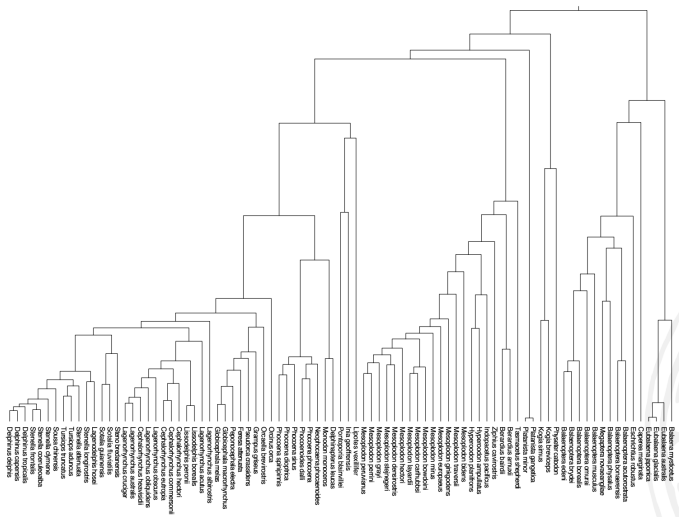
Such analyses are important for a wide range of life science applications.

The research front in many disciplines is partly defined today by our ability to learn the parameters of realistic phylogenetic models.

---

Recent models allow **diversification rates to vary across lineages**.

Such models can accommodate **diversification processes that gradually change over time**.



Fredrik Ronquist, Jan Kudlicka, Viktor Senderov, Johannes Borgström, Nicolas Lartillot, Daniel Lundén, Lawrence Murray, TS and David Broman. **Probabilistic programming: a powerful new approach to statistical phylogenetics**. June, 2020.

# Probabilistic programming

Has the potential to be **disruptive for the sciences** by providing rapid automated prototyping of algorithms and models for data.

**Probabilistic programming** provides a language with:

- support for declaring random variables,
- support for conditioning on the observed data,
- automatic inference.

Creates a clear **separation** between the model and the inference methods, encouraging model based thinking. Automate inference!

Our **first automated algorithm** – the Rao-Blackwellized particle filter. Massive potential for more automation!

`birch-lang.org`

Lawrence Murray and TS. **Automated learning with a probabilistic programming language: Birch**. *Annual Reviews in Control*, 46:29–43, 2018.

Lawrence M. Murray, Daniel Lundén, Jan Kudlicka, David Broman and TS. **Delayed sampling and automatic Rao-Blackwellization of probabilistic programs**. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, Lanzarote, Spain, April,

# AI 4 Research – new university-wide AI-lab

At Uppsala University we will **develop and make use of AI/ML for research**.

A **time-limited five year effort** consisting of an **antidisciplinary entity** from the entire university.

Located in newly refurbished premises at our main library Carolina Rediviva.

Key mechanism: **Internal AI sabbatical periods**

- Probably funded 50% by the entity and the rest by the department where the fellow remains employed/external grants.
- Duration: around 12 months.
- The fellows brings along 1-2 of their PhD students/post-docs.

# Deep probabilistic regression

# Background: regression using deep neural networks

**Supervised regression:** learn to predict a continuous output (target) value $y^\star \in \mathcal{Y} = \mathbb{R}^K$ from a corresponding input $x^\star \in \mathcal{X}$, given a training set $\mathcal{D}$ of i.i.d. input-output data

$$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N, \qquad (x_n, y_n) \sim p(x, y).$$

**Deep neural network (DNN):** a function $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \in \mathbb{R}^P$, that maps an input $x \in \mathcal{X}$ to an output $f_\theta(x) \in \mathcal{Y}$.

# Our ongoing work on deep regression

Deep learning for classification is handled using standard losses and output representations.

This is **not** the case when it comes to regression.

Train a model $p(y \mid x; \theta)$ of the conditional target density using a DNN to predict the un-normalized density **directly** from input-output pair $(x, y)$.

# Four existing approaches: 1. Direct regression

Train a DNN $f_\theta : \mathcal{X} \to \mathcal{Y}$ to directly predict the target $y^\star = f_\theta(x^\star)$.

Learn the parameters $\theta$ by minimizing a loss function $\ell(f_\theta(x_n), y_n)$, penalizing discrepancy between prediction $f_\theta(x_n)$ and ground truth $y_n$

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \ell(f_\theta(x_n), y_n), \qquad \theta = \arg\min_{\theta'} J(\theta').$$

Common choices for $\ell$ are the $L^2$ loss, $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$, and the $L^1$ loss.

Minimizing $J(\theta)$ then corresponds to minimizing the negative log-likelihood $\sum_{n=1}^{N} -\log p(y_n \mid x_n; \theta)$, **for a specific model** $p(y \mid x; \theta)$ of the conditional target density.

---

**Ex:** The $L^2$ loss corresponds to a fixed-variance Gaussian model:

$$p(y \mid x; \theta) = \mathcal{N}(y; f_\theta(x), \sigma^2).$$

# Four existing approaches: 2. Probabilistic regression

Why not explicitly employ this probabilistic perspective and try to create more **flexible** models $p(y \mid x; \theta)$ of the conditional target density $p(y \mid x)$?

**Probabilistic regression:** train a DNN $f_\theta : \mathcal{X} \to \mathcal{Y}$ to predict the parameters $\phi$ of a certain family of probability distributions $p(y; \phi)$, then model $p(y \mid x)$ with

$$p(y \mid x; \theta) = p(y; \phi(x)), \qquad \phi(x) = f_\theta(x).$$

The parameters $\theta$ are learned by minimizing $\sum_{n=1}^{N} -\log p(y_n \mid x_n; \theta)$.

---

**Ex:** A general 1D Gaussian model can be realized as:

$$p(y \mid x; \theta) = \mathcal{N}\big(y; \mu_\theta(x), \sigma^2_\theta(x)\big), \qquad f_\theta(x) = \begin{pmatrix} \mu_\theta(x) & \log \sigma^2_\theta(x) \end{pmatrix}^\mathsf{T} \in \mathbb{R}^2.$$

# Four existing approaches

The quest for improved regression accuracy has also led to the development of more specialized methods.

**3. Confidence-based regression:** train a DNN $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar confidence value $f_\theta(x, y)$, and maximize this quantity over $y$ to predict the target

$$y^\star = \arg\max_y f_\theta(x^\star, y)$$

---

**4. Regression-by-classification** Discretize the output space $\mathcal{Y}$ into a finite set of $C$ classes and use standard classification techniques.

# Background – Energy-based models (EBM)

An **energy-based models (EBM)** specifies a probability density

$$p(x;\theta) = \frac{e^{f_\theta(x)}}{Z(\theta)}, \qquad Z(\theta) = \int e^{f_\theta(x)} dx$$

directly via a parameterized scalar function $f_\theta(x)$.

By defining $f_\theta(x)$ using a **deep neural network**, $p(x;\theta)$ becomes expressive enough to learn practically any density from observed data.

---

EBMs have a rich history in machine learning:


LeCun, Y., Chopra, S., Hadsell, R. Ranzato, M and Huang, F. J. **A tutorial on energy-based learning**. In *Predicting structured data*, 2006.

Teh, Y. W., Welling, M., Osindero, S. and Hinton, G. E. **Energy-based models for sparse overcomplete representations**. *Journal of Machine Learning Research*, 4:1235–1260, 2003.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. **A neural probabilistic language model**. *Journal of machine learning research*, 3:1137–1155, 2003.

Hinton, G., Osindero, S., Welling, M. and Teh, Y-W. **Unsupervised discovery of nonlinear structure using contrastive backpropagation**. *Cognitive science*, 30(4):725–731, 2006.

Mnih, A. and Hinton, G. **Learning nonlinear constraints with contrastive backpropagation**. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2005.

Osadchy, M., Miller, M. L. and LeCun, Y. **Synergistic face detection and pose estimation with energy-based models**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2005.

# High-level description of our idea

**Confidence-based regression** give impressive results, but:

1. they require important (and tricky) task-dependent design choices
2. and usually lack a clear probabilistic interpretation.

**Probabilistic regression** is straightforward and generally applicable, but:

1. it can usually not compete in terms of regression accuracy.

Our construction **combines the benefits** of these two approaches while **removing the problems** above.

# Our (simple and very general) construction

A general regression method with a **clear probabilistic interpretation** in the sense that we learn a model $p(y \mid x, \theta)$ **without** requiring $p(y \mid x, \theta)$ to belong to a particular family of distributions.

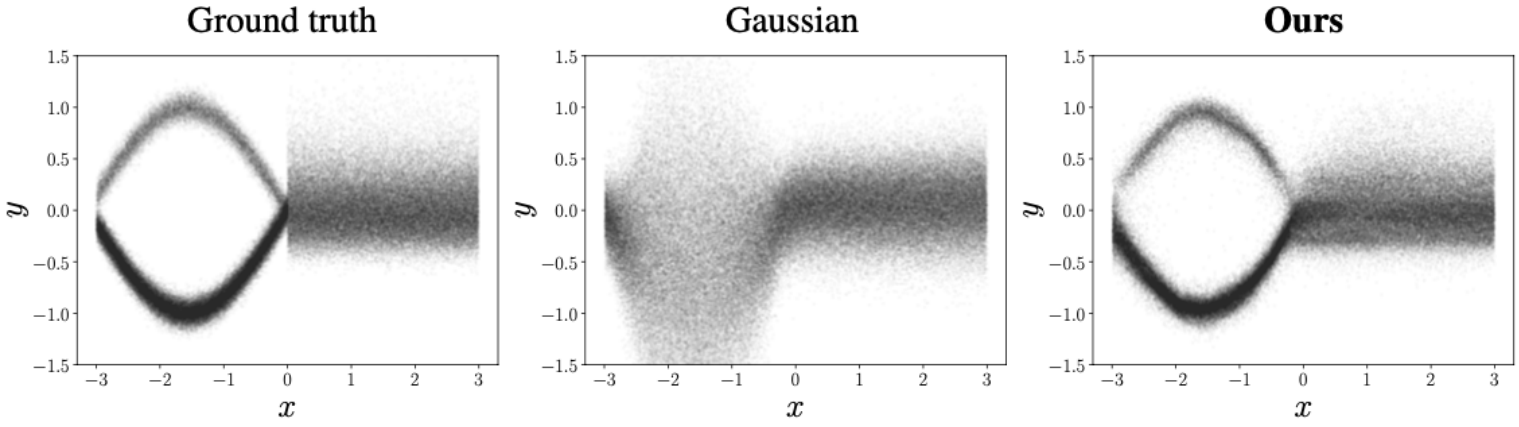
Let the DNN be a function $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that maps an input-output pair $\{x_n, y_n\}$ to a scalar value $f_\theta(x_n, y_n) \in \mathbb{R}$.

Define the resulting (flexible) probabilistic model as

$$p(y \mid x, \theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \qquad Z(x,\theta) = \int e^{f_\theta(x,y)} dy$$

# Learning flexible deep conditional target densities

1D toy illustration showing that we can learn multi-modal and asymmetric distributions, i.e. our model is **flexible**.




---

One good and one less good thing with our EBM:

**Good:** Its **modeling capacity** makes it highly attractive

**Less good:** Training is challenging.

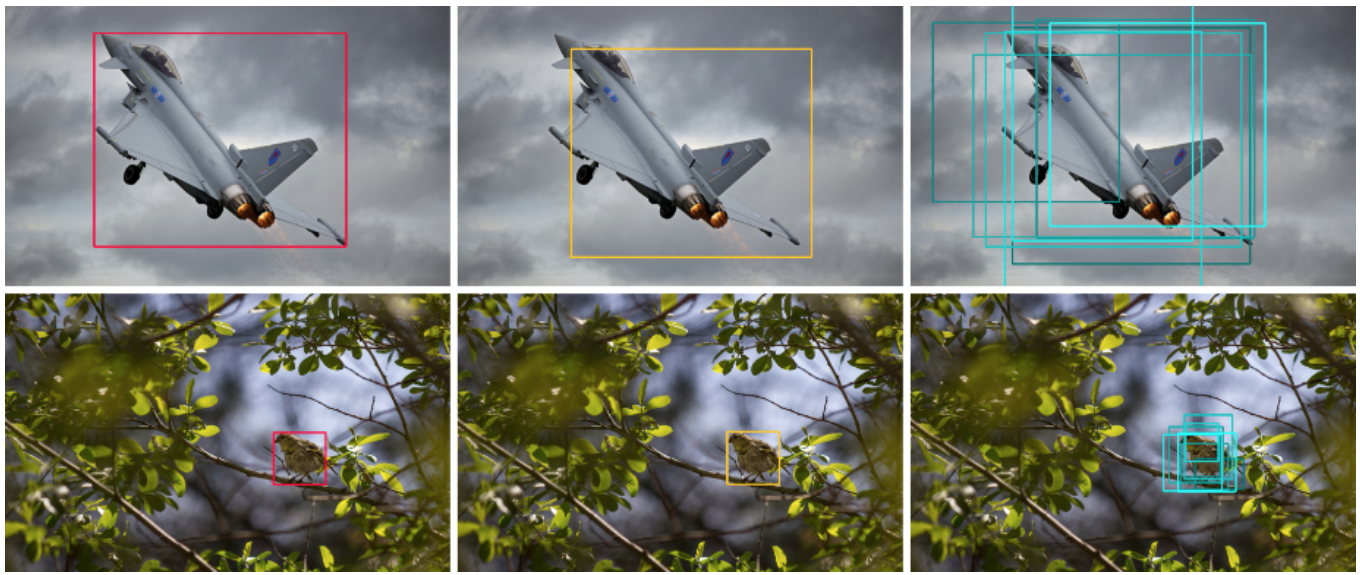
# Training the model

Simple, highly effective extension of noise contrastive estimation (NCE)

Michael Gutmann and Aapo Hyvärinen. **Noise-contrastive estimation: A new estima-tion principle for unnormalized statistical models**. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 297–304, 2010.

to train EBMs $p(y \mid x, \theta)$ for regression tasks.

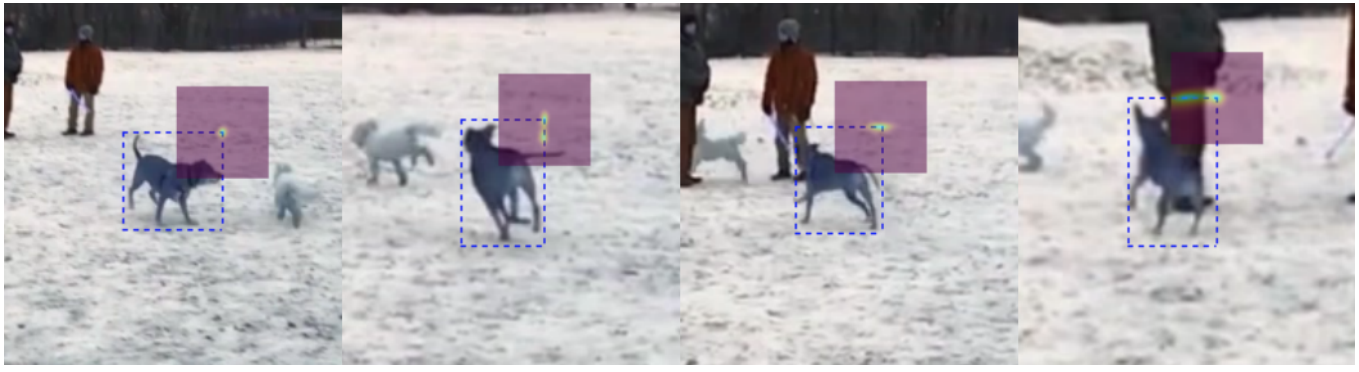
Our method can be understood as a direct generalization of NCE, **accounting for noise in the annotation process**.



# Experiments

Good results on four different computer vision (regression) problems:
1. Object detection, 2. Age estimation, 3. Head-pose estimation and
4. **Visual tracking**.

**Task (visual tracking):** Estimate a bounding box of a target object in every frame of a video. The target object is defined by a given box in the first video frame.



---


Fredrik K. Gustafsson, Martin Danelljan, Radu Timofte, TS. **How to train your energy-based model for regression**. April, 2020.

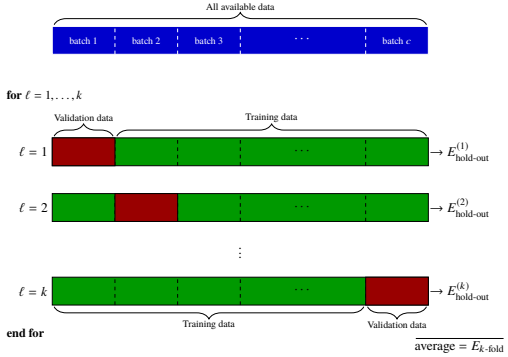
Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat and TS. **Learning deep conditional target densities for accurate regression**. November, 2019.

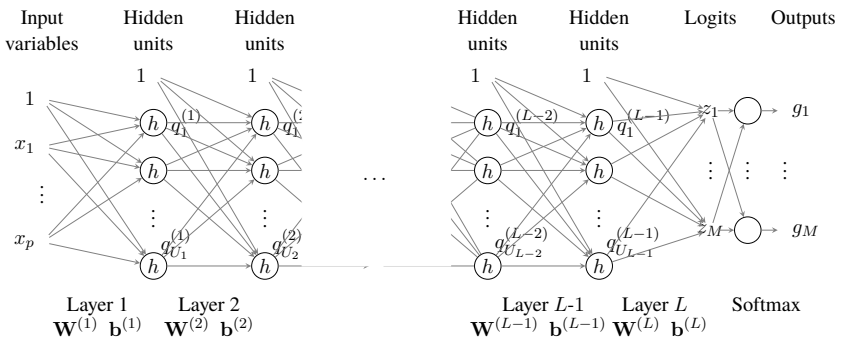
# Machine Learning education

# Machine Learning education – two new courses

**Statistical machine learning**: 5 hp for civilingenjörsstudenter year 4. **Supervised machine learning**, **from basic to state-of-the-art methods** (boosting, deep learning, . . . ). Mini-project + computer lab + written exam. Given since 2017, very popular among students.

www.it.uu.se/edu/course/homepage/sml/




---

Ongoing **book project** (with Cambridge University Press) for this course. Freely available (also after it is published in the beginning of 2021)

smlbook.org

# Machine Learning education – two new courses

**Advanced probabilistic machine learning**: 5 hp for civilingenjörsstudenter year 5. A **Bayesian perspective**, including **graphical models**, **approximate inference** and **Gaussian processes**, as well as **variational autoencoders**. Mini-project + computer lab + oral exam. Given for the first time in 2019.

www.it.uu.se/edu/course/homepage/apml/

$$
\begin{aligned}
\nabla_\xi \mathcal{L}_{\xi\zeta}(\mathbf{x}) &= \nabla_\xi \mathbb{E}_{q_\zeta(\mathbf{z}\,|\,\mathbf{x})}[\log p_\xi(\mathbf{x},\mathbf{z}) - \log p_\zeta(\mathbf{z}\,|\,\mathbf{x})] \\
&= \nabla_\xi \int (\log p_\xi(\mathbf{x},\mathbf{z}) - \log p_\zeta(\mathbf{z}\,|\,\mathbf{x}))\, q_\zeta(\mathbf{z}\,|\,\mathbf{x}) d\mathbf{z} \\
&= \mathbb{E}_{q_\zeta(\mathbf{z}\,|\,\mathbf{x})}[\nabla_\xi \log p_\xi(\mathbf{x},\mathbf{z})] \\
&\approx \nabla_\xi \log p_\xi(\mathbf{x},\mathbf{z}),
\end{aligned}
$$

Feel free to use these courses and our material!

# Conclusion

While ML techniques are used more and more in industry, scientists are—for good reasons—becoming aware of the potential in using ML in fundamental research.

The best predictive performance is currently obtained from **highly flexible learning systems**.

- Showed three examples of **AI for Research**.
- Introduced a useful model and learning algorithm for deep regression.

Remember to talk to people who work on **different problems** with **different tools**!! (Visit other fields!)

**Trevlig midsommar!!!**