UPPSALA
UNIVERSITET

Machine Learning - trends and tools

Thomas Schön
Uppsala University

SEB Data Science Community
April 25, 2019.

"Machine learning gives computers the ability to **learn without being explicitly programmed** *for the task at hand."*

"Anyone making confident predictions about anything having to do with the future of artificial intelligence is either kidding you or kidding themselves."

Andrew McAfee, MIT

What we do in the team — who we are

We automate the extraction of knowledge and understanding from data.

Both basic research **and** applied research (with companies).




Create **probabilistic models** of dynamical systems and their surroundings.

Develop methods to **learn** models from data.

The models can then be used by machines (or humans) to **understand** or **make decisions** about what will happen next.

What do I hope to achieve today?

1. Briefly introduce the scientific field of Machine Learning.
2. Create an **awareness/interest** around this technology.
3. **A bit more specific:** Give a few concrete examples offering a (hopefully) intuitive understanding.

What is machine learning all about?

Machine learning is about learning, reasoning and acting based on data.

Machine learning gives computers the ability to **learn without being explicitly programmed** for the task at hand.

"It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science."

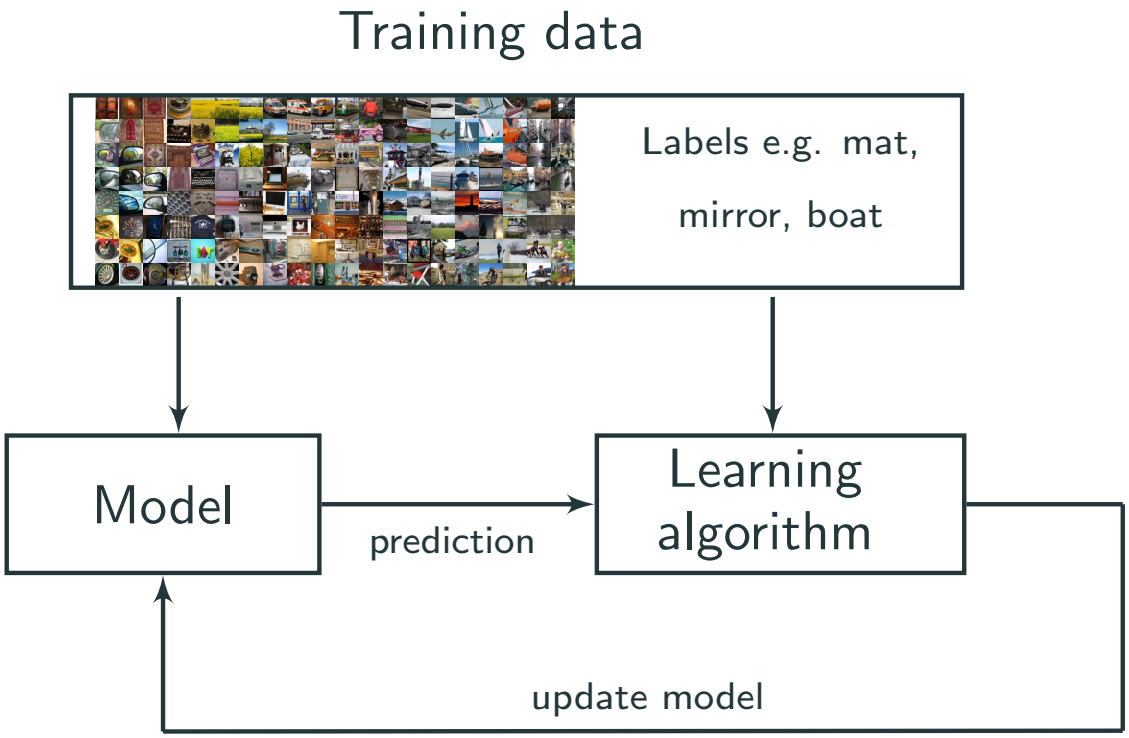
Ghahramani, Z. **Probabilistic machine learning and artificial intelligence**. *Nature* 521:452-459, 2015.

Jordan, M. I. and Mitchell, T. M. **Machine Learning: Trends, perspectives and prospects**. *Science*, 349(6245):255-260, 2015.

Machine Learning (supervised)

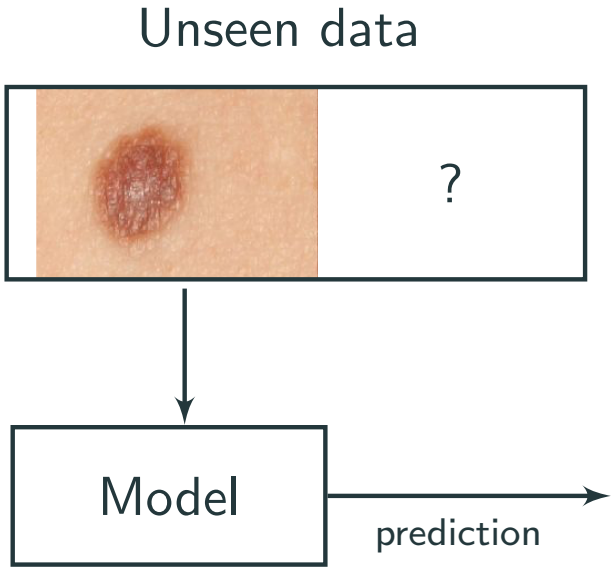
Data on its own is typically useless, it is only when we can extract knowledge from the data that it becomes useful.

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.

Unsupervised, reinforcement and semi-supervised learning.

The four cornerstones

Cornerstone 1 **(Data)** Typically we need lots of it.

Cornerstone 2 **(Mathematical model)** A mathematical model is a compact representation of the data that in precise mathematical form captures the key properties of the underlying situation.

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

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

Mathematical models – representations

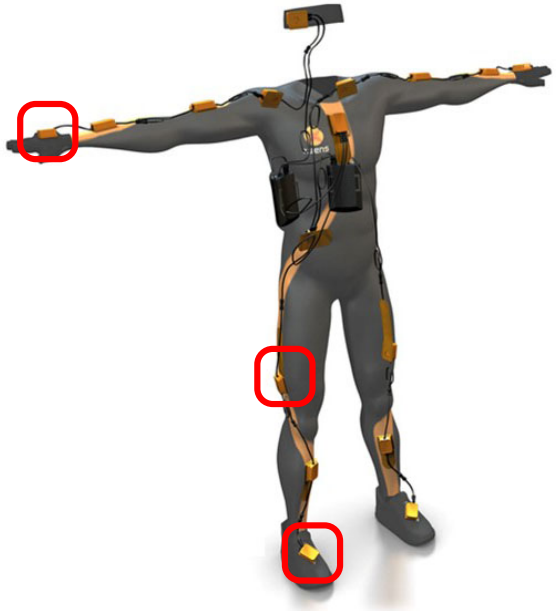
The performance of an algorithms typically depends on which representation (model) that is used for the data.

When solving a problem – start by thinking about **which model/representation to use**!

International Conference on Learning Representations (ICLR)
www.iclr.cc/

Ex (Classical Engineering) – Motion estimation

Aim: Compute the position and orientation of the different body segments of a person moving around indoors (motion capture).



What is a mathematical model?

Illustrate the use of three different models:

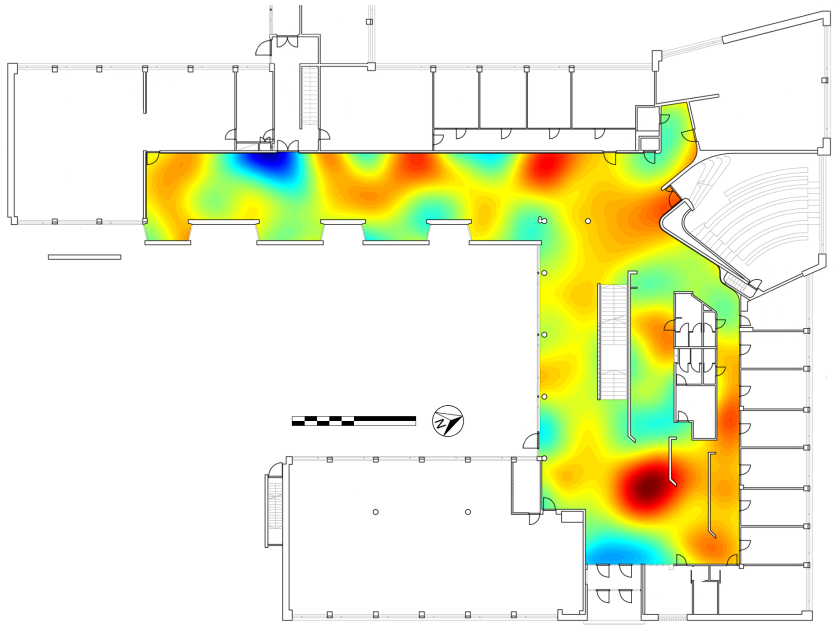
1. Integration of sensor observations.
2. Add a biomechanical model.
3. Add a world model.

Ex (Machine Learning) – 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.

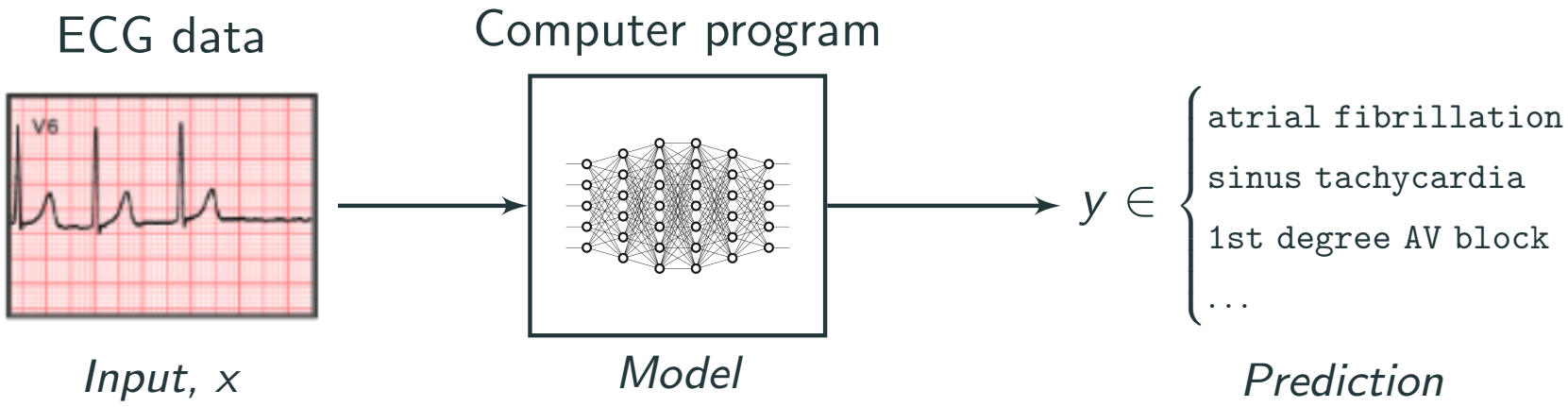


www.youtube.com/watch?v=enlMiUqPVJo

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.

Ex (Machine Learning) – Automatic ECG classification



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.

The model – learning relationship

The problem of learning (estimating) a model based on data leads to computational challenges, both

- **Integration:** e.g. the HD integrals arising during marg. (averaging over all possible parameter values z):

$$p(D) = \int p(D \mid z) p(z) \mathrm{d}z.$$

- **Optimization:** e.g. when extracting point estimates, for example by maximizing the posterior or the likelihood

$$\widehat{z} = \arg\max_{z} \; p(D \mid z)$$

Typically impossible to compute exactly, use approximate methods

- Monte Carlo (MC), Markov chain MC (MCMC), and sequential MC (SMC).
- Variational inference (VI).
- Stochastic optimization.

Key lesson from contemporary Machine Learning

Flexible models often give the best 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. **Probabilistic machine learning and artificial intelligence**. *Nature* 521:452-459, 2015.

"With enough training data the machine can be trained to make very good predictions from previously unseen data."

Outline

1. What is machine learning?
2. Models – a few examples
3. **Flexible models**
   a) **Deep learning**
   b) **Gaussian processes**
4. Short overview of our research topics (if there is time)
5. Conclusion

Machine learning gives computers the ability to **learn without being explicitly programmed** for the task at hand.

Deep learning – what is it?

The mathematical model has been around for 70 years, but over the last $5-7$ years 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

The underlying model is a big mathematical function with **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 θ.

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 θ is the parameters composed by the "weights" θ_i and the offset ("bias") term θ_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.

Deep neural networks

Let the computer **learn from experience** and understand the situation in terms of a **hierarchy of concepts**, where each concepts is defined in terms of its relation to simpler concepts.

If we draw a graph showing these concepts of top of each other, the graph is **deep**, hence the name deep learning.

It is accomplished by using **multiple levels of representation**. Each level transforms the representation at the previous level into a new and more abstract representation,

$$z^{(l+1)} = f\left(\Theta^{(l+1)} z^{(l)} + \theta_0^{(l+1)}\right),$$

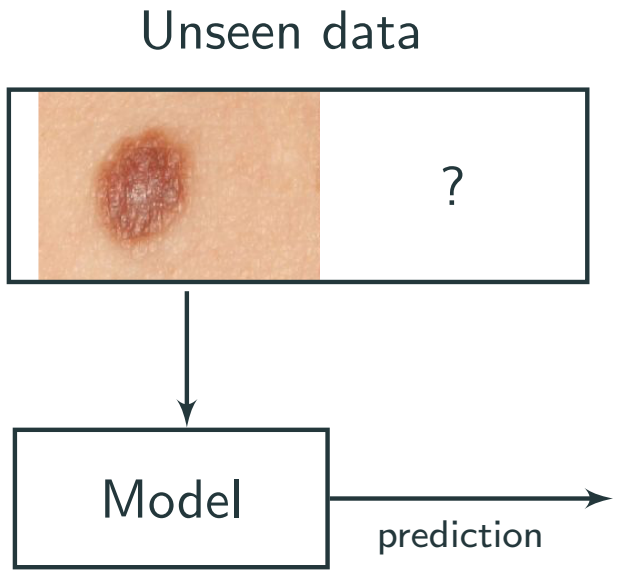
starting from the input (raw data) $z^{(0)} = x$.

Key aspect: The layers are **not** designed by human engineers, they are generated from (typically lots of) data using a learning procedure and lots of computations.

Deep learning – example (skin cancer)

Start from a mathematical model trained on 1.28 million images (**transfer learning**). Make minor modifications of it, specializing to present situation.

Learn new model parameters using 129 450 clinical images ($\sim$ 100 times more images than any previous study).

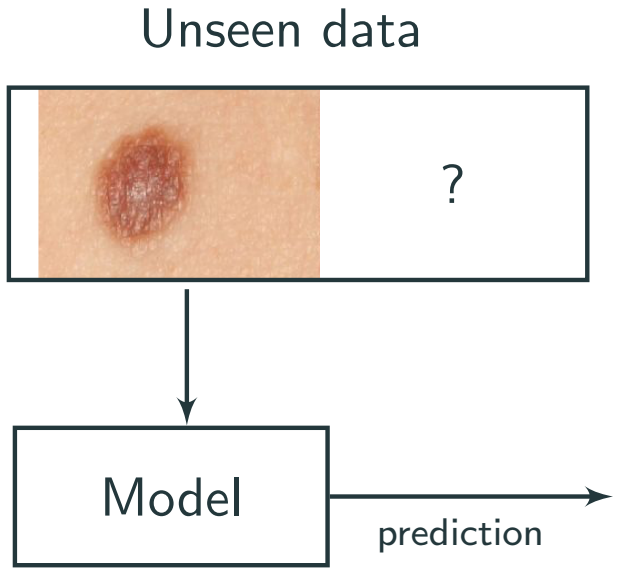


Andre Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S. **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature*, 542, 115–118, February, 2017.

Deep learning – example (skin cancer)

Start from a mathematical model trained on 1.28 million images (**transfer learning**). Make minor modifications of it, specializing to present situation.

Learn new model parameters using 129 450 clinical images ($\sim$ 100 times more images than any previous study).

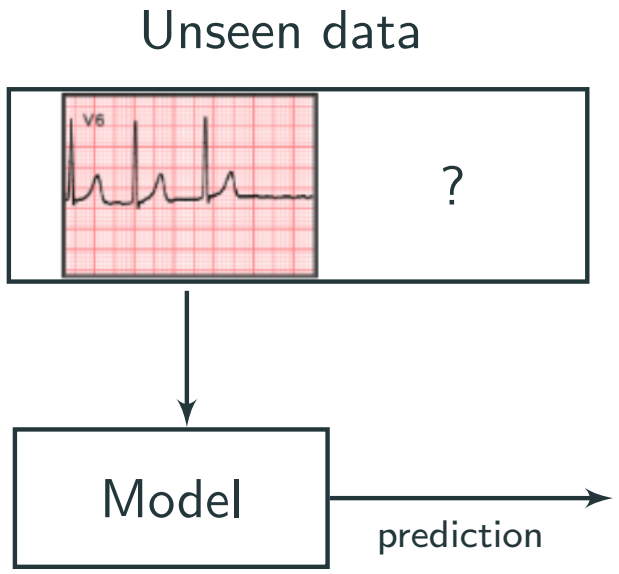


The results are on par with professional dermatologists on specific tasks. Still, far from being clinically useful, but at least they give us "valid reasons to remain cautiously optimistic" as someone said.

Andre Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S. **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature*, 542, 115–118, February, 2017.

ECG classification – the CODE study

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



Background: Joint work with medical doctors 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.

Ribeiro, A. H., Ribeiro, M. H., Paixao, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira, W., TS and Ribeiro, A. L. P. **Automatic diagnosis of the short-duration 12-lead ECG using a deep neural network: the CODE study**. Submitted, 2019.

Outline

1. What is machine learning?
2. Models – a few examples
3. Flexible models
   a) Deep learning
   b) Gaussian processes
4. Short overview of our research topics (if there is time)
5. Conclusion

Machine learning gives computers the ability to **learn without being explicitly programmed** for the task at hand.

The Gaussian process is a model for nonlinear functions

Q: Why is the Gaussian process used everywhere?

It is a **non-parametric** and **probabilistic** model for nonlinear functions.

- **Non-parametric** means that it does not rely on any particular parametric functional form to be postulated.
- **Probabilistic** means that it takes uncertainty into account in every aspect of the model.

An abstract idea

In probabilistic (Bayesian) linear regression

$$y_t = \underbrace{\beta^{\mathsf{T}} \mathbf{x}_t}_{f(\mathbf{x}_t)} + e_t, \qquad e_t \sim \mathcal{N}(0, \sigma^2),$$

we place a prior on β, e.g. $\beta \sim \mathcal{N}(0, \alpha^2 I)$.

(Abstract) idea: What if we instead place a prior directly on the function $f(\cdot)$

$$f \sim p(f)$$

and look for $p(f \mid y_{1:T})$ rather than $p(\beta \mid y_{1:T})$?!

One concrete construction

Well, one (arguably simple) idea on how we can reason probabilistically about an unknown function f is by assuming that $f(x)$ and $f(x')$ are jointly Gaussian distributed

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}(m, K).$$

If we accept the above idea we can without conceptual problems generalize to any *arbitrary* finite set of input values $\{x_1, x_2, \ldots, x_T\}$.

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_T) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \ldots & k(x_1, x_T) \\ \vdots & \ddots & \vdots \\ k(x_T, x_1) & \ldots & k(x_T, x_T) \end{pmatrix} \right)$$

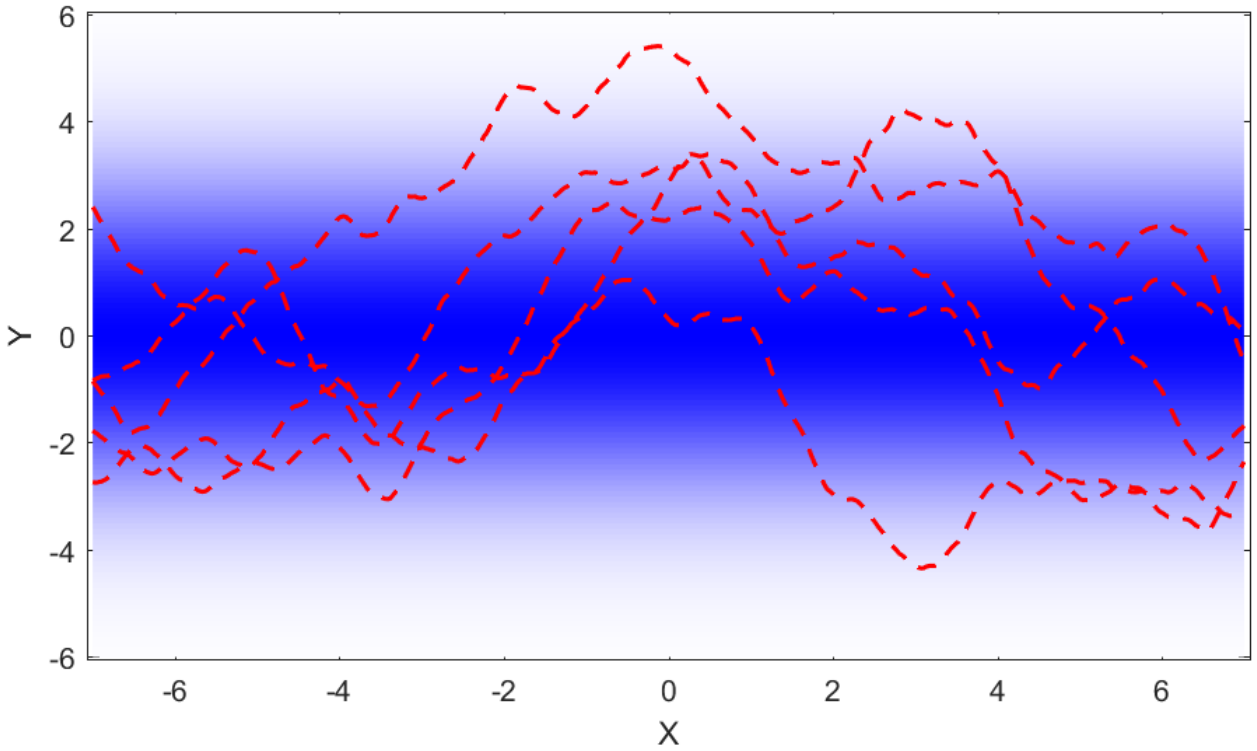
Definition

Definition: (Gaussian Process, GP) A GP is a (potentially infinite) collection of random variables such that any finite subset of it is jointly distributed according to a Gaussian.

We now have a prior!

$$f \sim \mathcal{GP}(m, k)$$

The GP is a **generative** model so let us first sample from the prior.



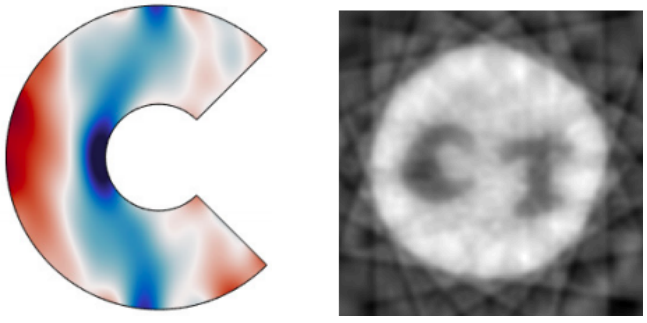
GP regression – illustration

Snapshot — Constrained GP for tomographic reconstruction

Tomographic reconstruction goal: Build a map of an unknown quantity within an object using information from irradiation experiments.

Ex1) Modelling and reconstruction of strain fields.

Ex2) Reconstructing the internal structure from limited x-ray projections.



Carl Jidling, Johannes Hendriks, Niklas Wahlström, Alexander Gregg, TS, Chris Wensrich and Adrian Wills. **Probabilistic modelling and reconstruction of strain**. *Nuclear inst. and methods in physics research: section B*, 436:141-155, 2018.

Zenith Purisha, Carl Jidling, Niklas Wahlström, Simo Särkkä and TS. **Probabilistic approach to limited-data computed tomography reconstruction**. *Draft*, 2019.

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

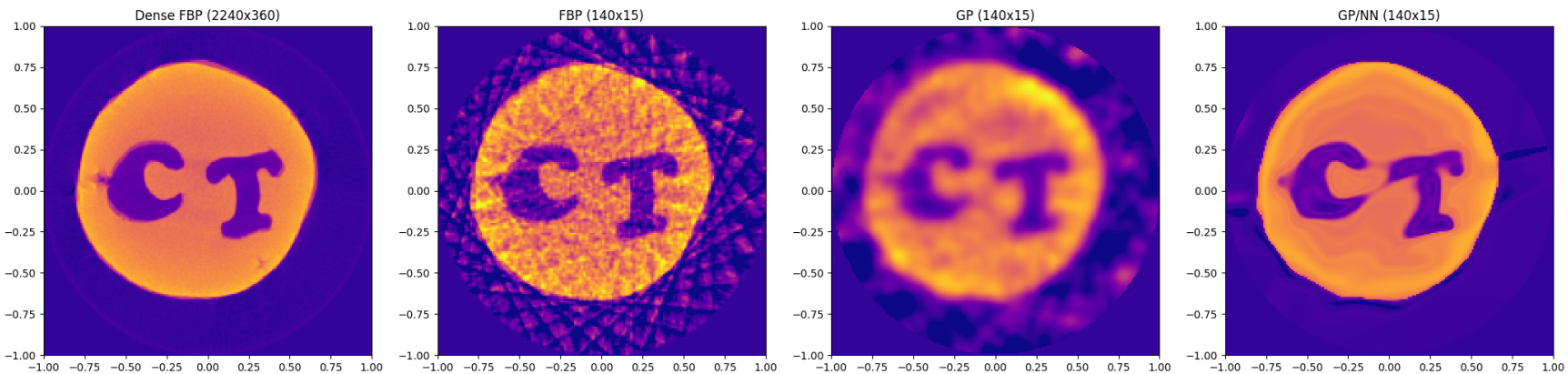
Ongoing work: We can also combine DL and GPs!

Problem: The standard stationary kernels can not deal with noon-smooth features, such as rapid, step-like changes.

Solution: Learn a deep NN transforming the inputs to a stationary kernel. Known as manifold Gaussian processes or deep kernel learning.

Our contribution: Developed a model that efficiently handles measured data that is expressed as line integrals of the unknown function.

Relevance: Tomographic reconstruction problems are on this form.



We are currently in the process of properly explaining this ...

Data Consistency Check

Question: Is a model class $p(y \mid \theta), \theta \in \Theta$ consistent (whatever that could mean . . .) with given data y?

Existing methods:

Compare models
Cross-validation, **AIC**, **BIC**, . . .

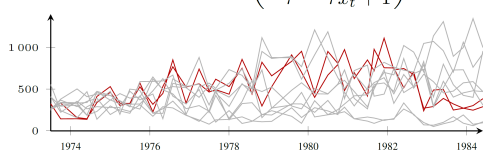
Validate a model
Ljung-Box (AR-like models)
Kolmogornov-Smirnov, **Anderson-Darling**, . . . (1-D IID data)
Very few (if any?) **general methods**

Our idea: Simulate data from the model and compare it (in terms of likelihoods) to the given data.

$\rightarrow$ A very general (but computer-intensive) approach!

Kangaroo count model

$$x_1 = \mathcal{LN}(0, 5),$$
$$\frac{dx_t}{x_t} = \frac{\sigma^2}{2}dt + \sigma dW_t,$$
$$y_{1,t}, y_{2,t} \mid x_t \overset{\text{i.i.d.}}{\sim} \mathcal{NB}\left(\frac{\tau - 1}{\tau}, \frac{\tau x_t}{\tau x_t + 1}\right),$$



Conclusion: Model not inconsistent with observed data!

Earthquake count models

Magnitude	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
≥ 8	0	1	1	0	1	2	1	2	4	0	1	1	1	2
≥ 7	18	15	16	13	15	16	11	11	18	12	17	24	20	16
≥ 6	136	160	137	139	156	157	151	153	196	179	161	175	207	133
≥ 5	1192	1495	1352	1309	1364	1672	1843	1877	2283	1965	2075	2395	2692	1680

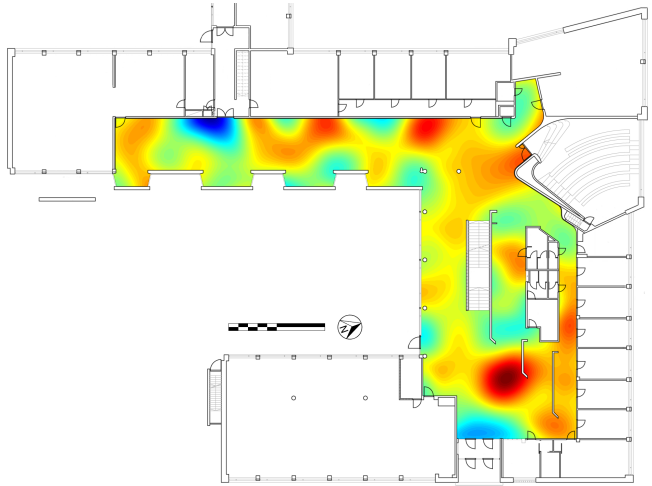
Magnitude	Poisson distribution	Negative binomial distribution
≥ 8	$\hat{\text{PFA}}^\star = 0.40$	$\hat{\text{PFA}}^\star = 0.39$
≥ 7	$\hat{\text{PFA}}^\star = 0.29$	$\hat{\text{PFA}}^\star = 0.38$
≥ 6	$\hat{\text{PFA}}^\star = 0.00$	$\hat{\text{PFA}}^\star = 0.30$
≥ 5	$\hat{\text{PFA}}^\star = 0.00$	$\hat{\text{PFA}}^\star = 0.13$

Conclusion: Poisson model inconsistent with "small" earthquakes!

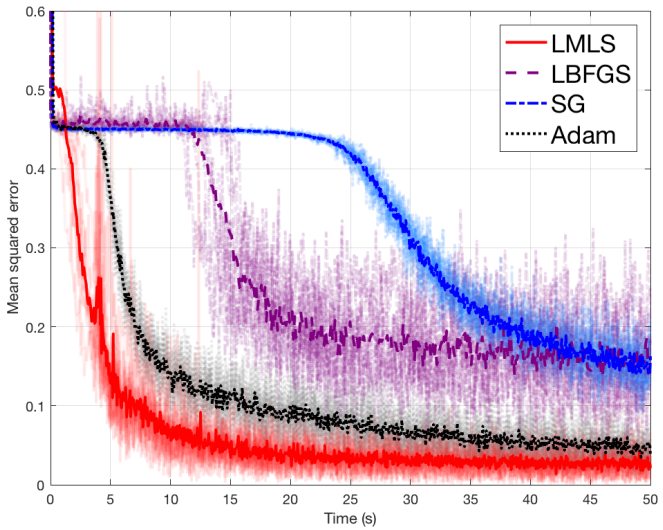
Andreas Lindholm, Dave Zachariah, Petre Stocia and TS. **Data Consistency Approach to Model Validation.** *IEEE Access*, 2019. (accepted this morning). *arXiv:1808.05889*. `github.com/saerdna-se/consistency-criterion`

Active research topics – slightly technical

1. **Probabilistic modelling**
   a) General: Flexible models, in particular the Gaussian process (GP), deep GPs.
   b) Specific: **Dynamical** phenomena and their surroundings.

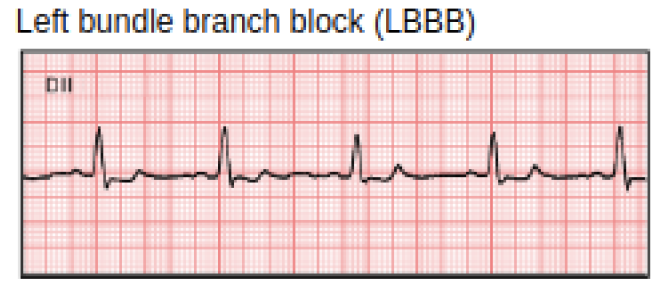


2. **New relevant algorithms**
   a) Large-scale optimization
   b) Approximate integration/inference
      i) Sequential Monte Carlo
      ii) Variational inference
      iii) Markov chain Monte Carlo



Active research topics – slightly technical

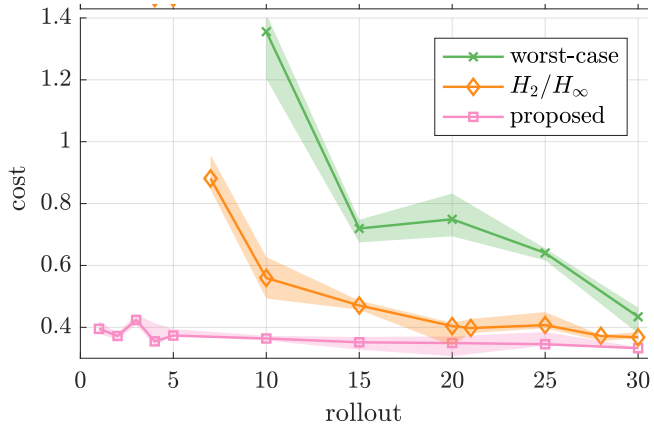
3. **Deep learning (DL)**
   a) Deep probabilistic constructions
   b) Representing (and understanding) uncertainty within deep learning (including Bayesian DL)
   c) Mathematical understanding of DL



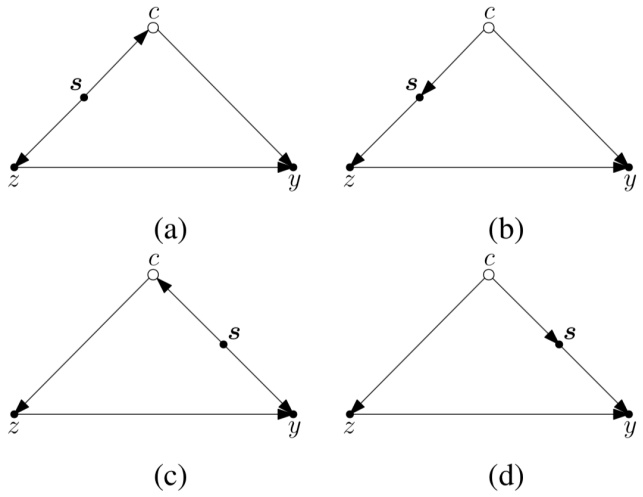
4. **Probabilistic programming**
   a) Potential to automate modelling!
   b) Developing our own probabilistic programming language (Birch)

Active research topics – slightly technical

5. **Reinforcement learning/control**
   a) Learn how to control
   b) Mathematical guarantees



6. **Causality** (new topic)
   a) Aiming to learn causal relationships (not just associations/correlations)
   b) Naturally leads to the need for combining human knowledge **and** data.



7. **Self-supervised** learning (new topic)
   a) Use small amount of labeled data and large amounts of unlabeled data.

Education – Our Machine Learning courses and new initiatives

PhD level courses in Machine Learning

1. Probabilistic Machine Learning (given since 2011) 50 students
2. Deep Learning (first time this spring) 50 students
3. Sequential Monte Carlo (given since 2012) 80 students

MSc level courses in Machine Learning

1. Statistical Machine Learning (given since 2017) 160 students
2. Probabilistic Machine Learning (first time this autumn)

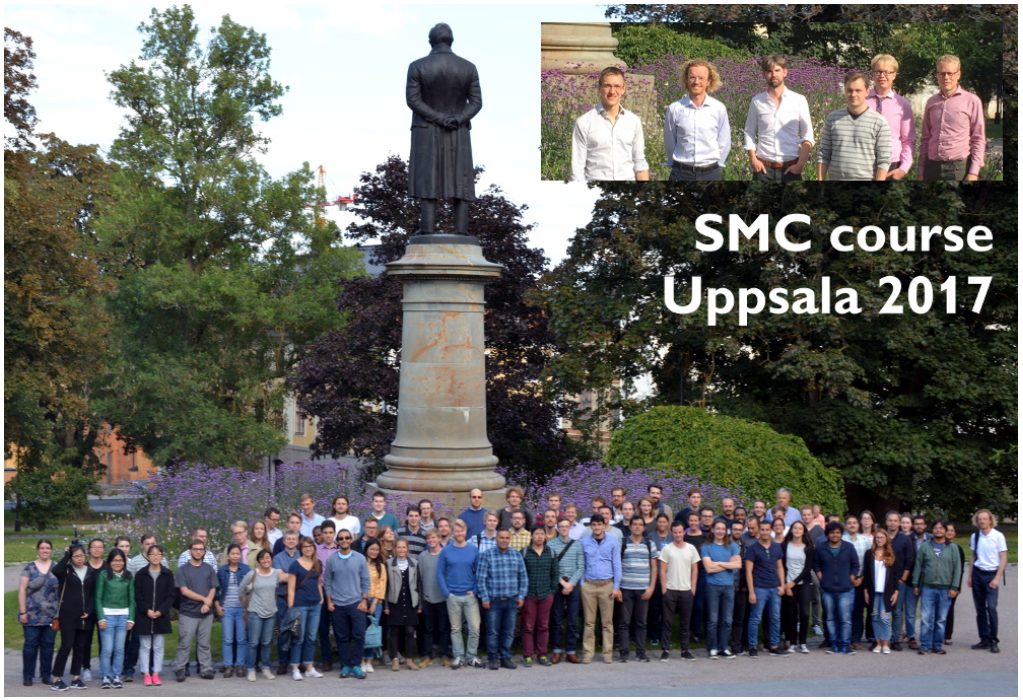
New MSc programs involving Machine Learning (starting in 2020)

1. Data Science
2. Image processing and Machine Learning

Intensive PhD level course on SMC in August

Intensive PhD course on **SMC methods** in August 2019.

Sequential Monte Carlo (SMC) is a random-sampling-based class of methods for approximate inference. Perfect for problems in nonlinear time series, but it is indeed much more generally applicable.



www.it.uu.se/research/systems_and_control/education/2019/smc

Christian A. Naesseth, Fredrik Lindsten and TS. *Elements of Sequential Monte Carlo*. *arXiv:1903.04797*, March, 2019.

What did I hope to achieve today?

1. Briefly introduce the scientific field of Machine Learning.
2. Create an **awareness/interest** around this technology.
3. **A bit more specific:** Give a few concrete examples offering a (hopefully) intuitive understanding.

Conclusion

Machine learning gives computers the ability to **learn without being explicitly programmed** for the task at hand.

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

We discussed two flexible model classes:

1. Deep learning
2. Gaussian processes

Uncertainty is a key concept!

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