UPPSALA
UNIVERSITET

Probabilistic modelling – driven by data, guided by physics

Thomas Schön
Uppsala University

Mathematics for Complex Data
June 24, 2019.

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

Machine Learning – the four cornerstones

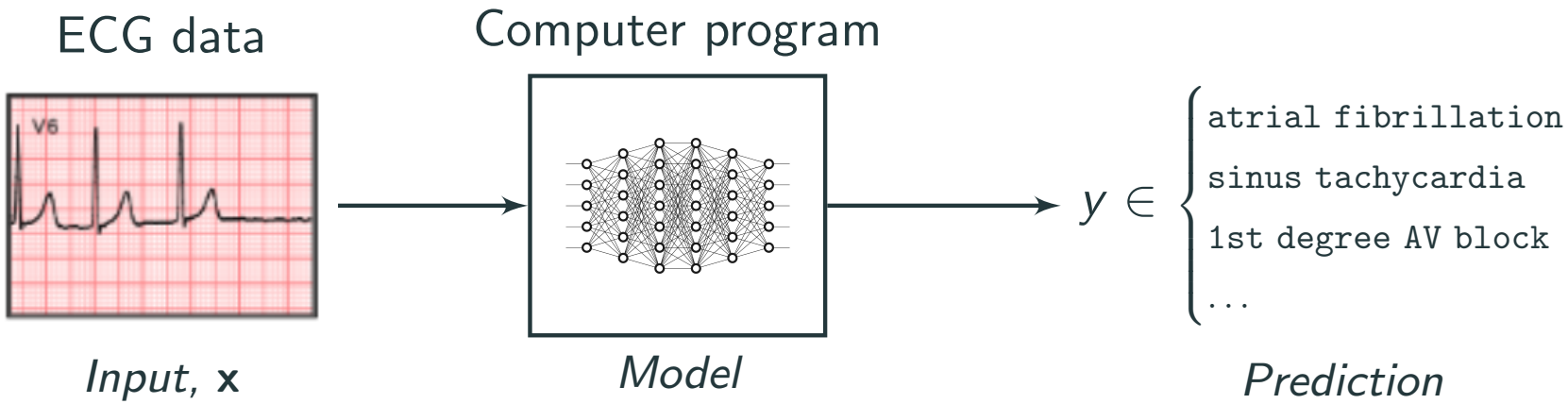
Cornerstone 1 **(Data)** Typically we need lots of it.

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.

Ex – Automatic ECG classification



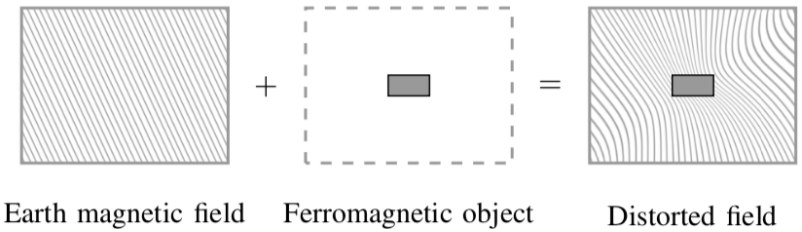
We are now reaching human level (medical doctor) performance on certain specific tasks.

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.

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**, *arXiv:1904.01949*, 2019.

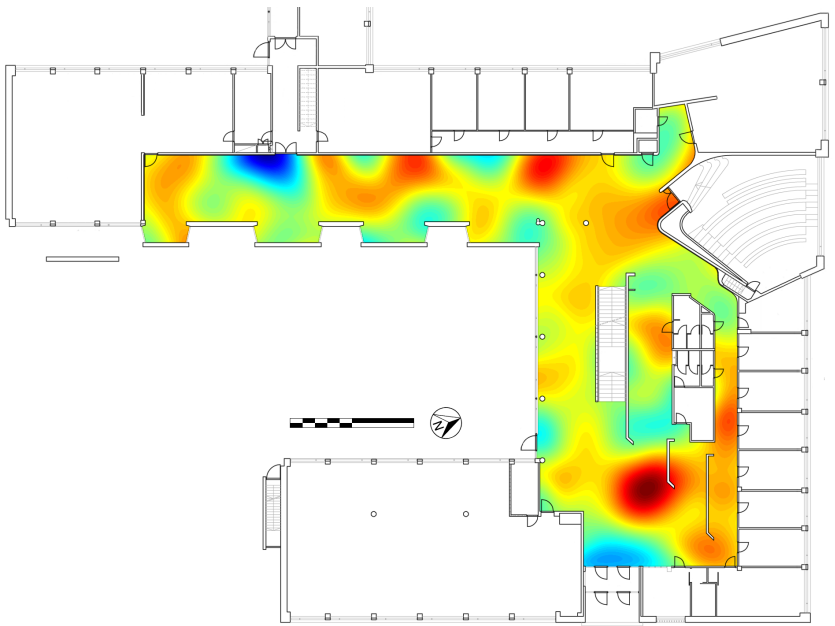
Ex – Ambient magnetic field map



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

Ex – Indoor localization using deviations in the magnetic field

Aim: Compute the **position** using variations in the ambient magnetic field and the motion of the person (acceleration and angular velocities). All of this observed using sensors in a standard smartphone.

Show movie!

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.

Arno Solin, Simo Särkkä, Juho Kannala and Esa Rahtu. **Terrain navigation in the magnetic landscape: Particle filtering for indoor positioning.** In *Proceedings of the European Navigation Conference*, Helsinki, Finland, June, 2016.

Probabilistic modelling – representation of beliefs (uncertainty)

For a machine to behave intelligently I believe it needs the

capability to represent and manipulate beliefs/uncertainty

about the real world.

As the machine perceives the world via its sensors it must then update its beliefs in light of the new information.

The mathematics of **probability theory** is well developed and

1. it allows us to not only **represent** uncertainty,
2. but it also prescribes how to **manipulate** it based on the information in new measurements.

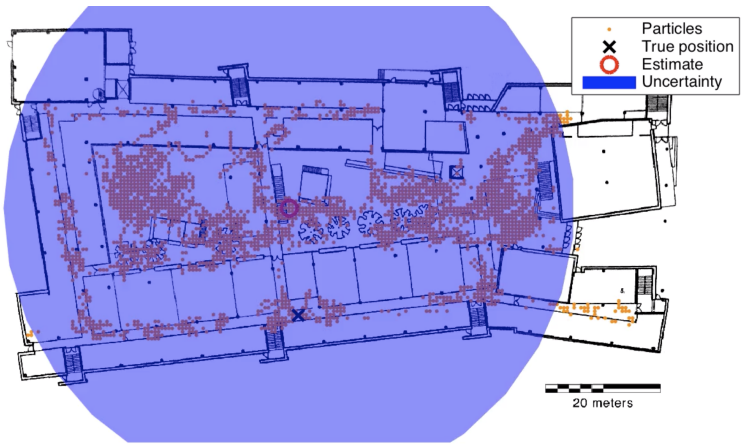
Probabilistic modelling – representation of beliefs (uncertainty)

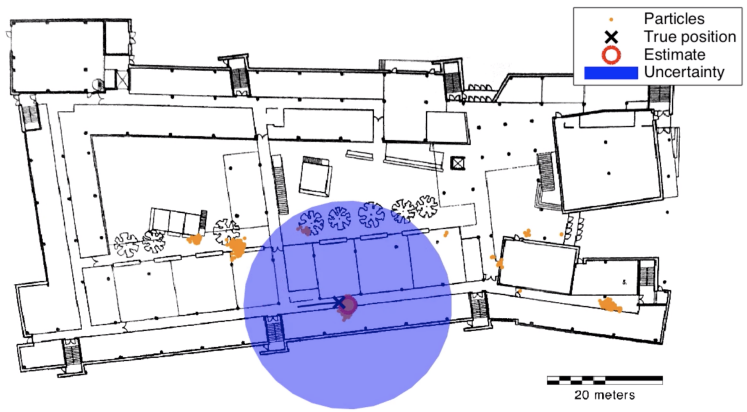
A very important fact is that **inverse probability** (i.e. Bayes rule)

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

allows us to infer unknown variables (x), adapt our models, make predictions and learn from data (y).

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





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.

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.

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 made in the simple combination of flexible data-driven models and solid widely available knowledge that we already have.

Aim of this talk: Try to provide some concrete evidence for my opinion (and to introduce the GP).

Blending prior knowledge and data

Resulting technical challenge: How can we use known structure 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?

Vision

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

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. Now, it is really time to become a bit more precise...

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

Outline

1. Introduction
2. Models – a few examples
3. Vision – blending prior knowledge and data
4. **A flexible model – the Gaussian process**
5. GPs with line integral measurements
6. GP + deep learning with integral measurements

(7. Strain field reconstruction from neutron diffraction experiments)

8. 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 $p(\beta \mid y_{1:T})$?!

$y_{1:T} = \{y_1, \ldots, y_T\}$

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 $\{x_1, x_2, \ldots, x_T\}$.

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

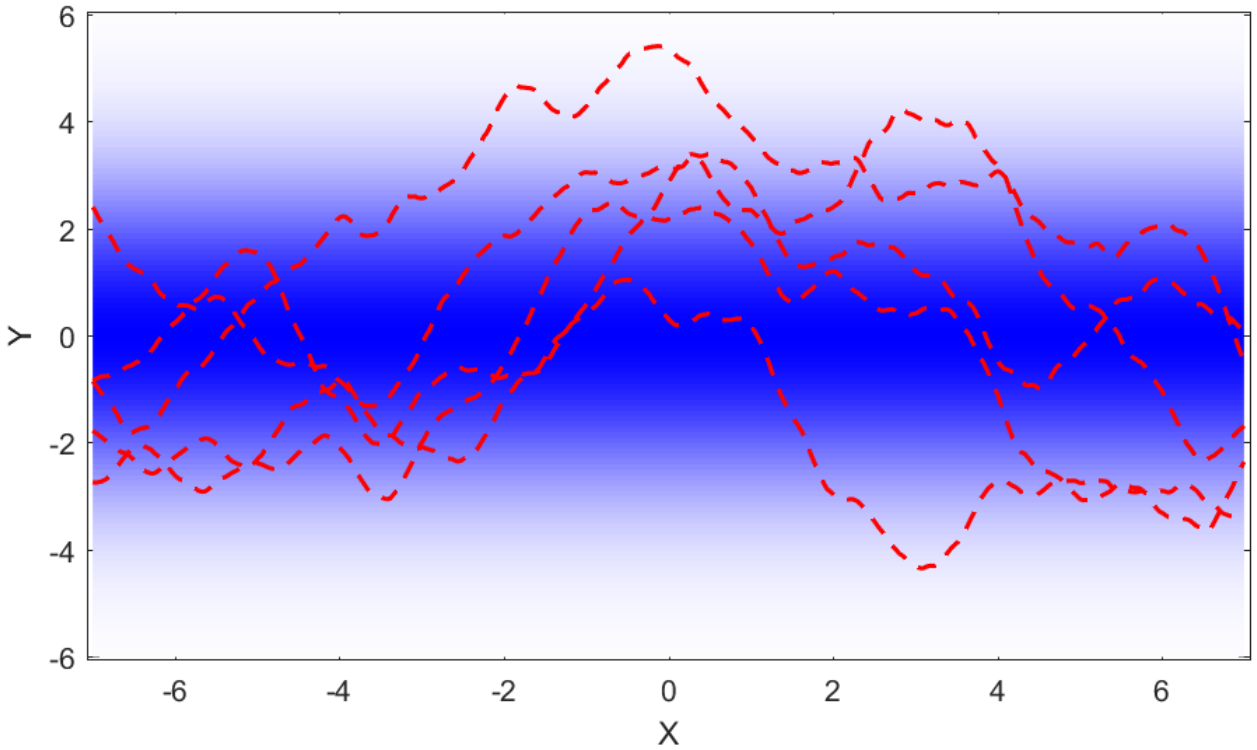
Definition

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.

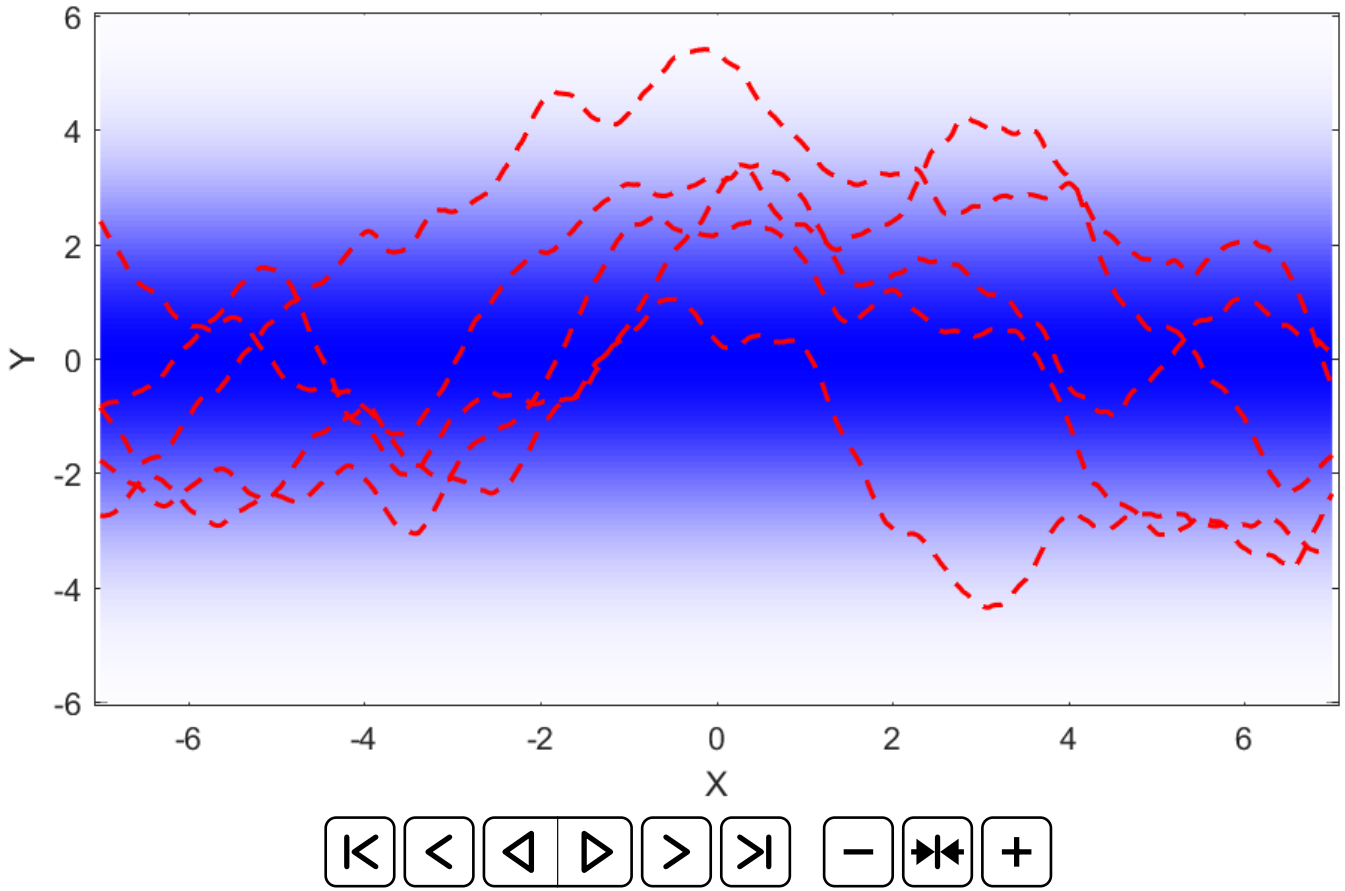
We now have a prior!

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

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



GP regression – illustration



Starting somewhere – more concrete from hereon

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

2	16	13	3
11	5	8	10
7	9	12	6
14	4	1	15

4	9	2
3	5	7
8	1	6

Computed tomography (CT)

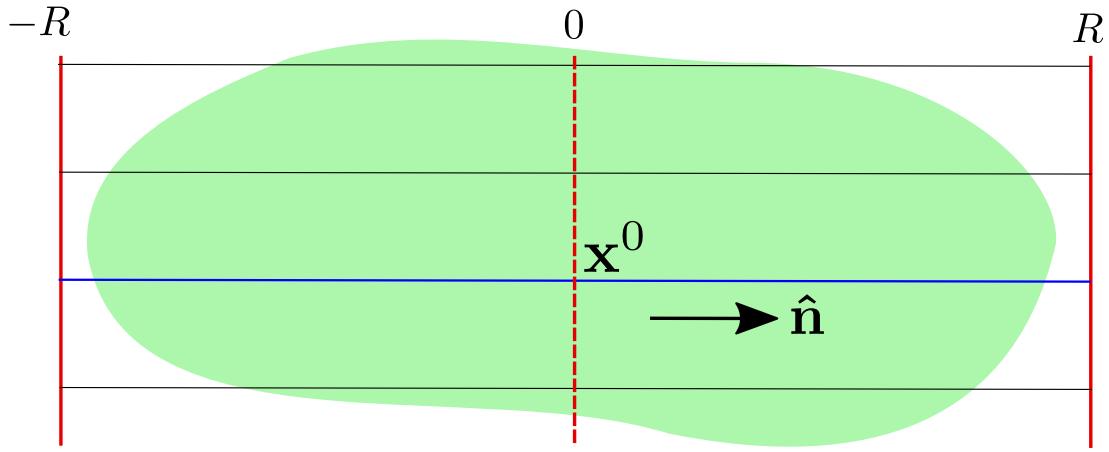
Tomographic reconstruction: Recover the internal structure

$$f(\mathbf{x}),\ \mathbf{x} = [x\ y]^{\mathsf{T}}$$

of an object from irradiation experiments.

Line integral measurements

$$y = \int_{-R}^{R} f(\mathbf{x}^0 + s\hat{\mathbf{n}})ds + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$



Limited data (sparse projections) important.

Linear functional measurements in GPs (more general)

Model the target function $f(\mathbf{x})$ as a GP

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

Fact: a GP is closed under linear transformations:

$$\mathcal{L}f(\mathbf{x}) \sim \mathcal{GP}(0, \mathcal{L}\mathcal{L}'k(\mathbf{x}, \mathbf{x}'))$$

where for us (in the CT case)

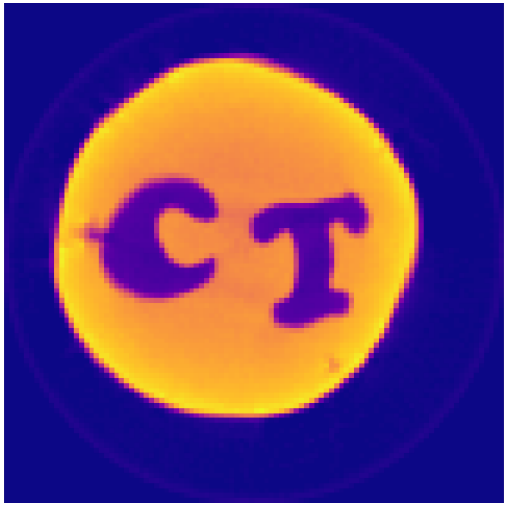
$$\mathcal{L}f(\mathbf{x}) = \int_{-r}^{r} f(\mathbf{x}^0 + s\hat{\mathbf{n}})ds,$$

Our CT and strain field reconstruction examples have measurements:

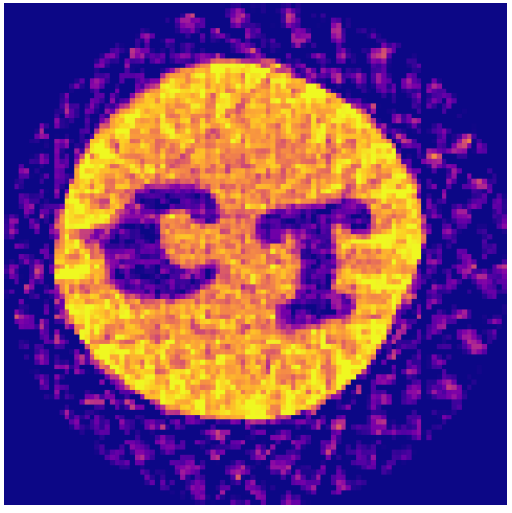
$$y = \int_{-r}^{r} f(\mathbf{x}^0 + s\hat{\mathbf{n}})ds + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, Q)$$

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.

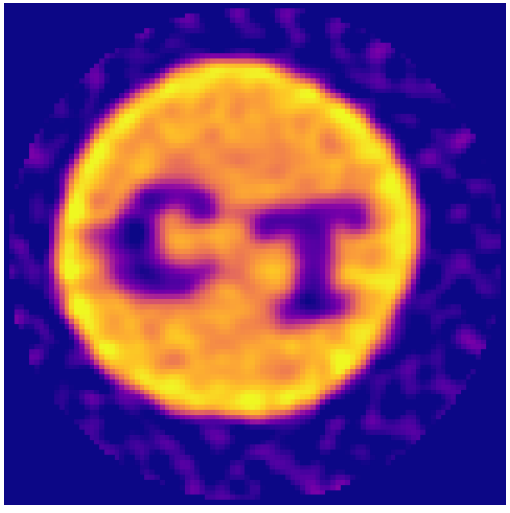
Ex. CT – carved cheese experiment



Ground truth



FBP



GP

Question: Why is the GP solution so blurry?

All details on this construction are available in

Zenith Purisha, Carl Jidling, Niklas Wahlström, Simo Särkkä, TS. **Probabilistic approach to limited-data computed tomography reconstruction**, *arXiv:1809.03779*, 2018.

Extending the expressiveness to non-stationary behaviors

The covariance function $k(\mathbf{x}, \mathbf{x}')$, stipulates the basic behavior of the target function $f(\mathbf{x})$.

The selection of $k(\mathbf{x}, \mathbf{x}')$ is the most crucial part of GP modelling.

Extend the expressiveness of stationary covariance functions by transforming the inputs through a nonlinear mapping $u(\cdot)$ to form $k(u(\mathbf{x}), u(\mathbf{x}'))$, effectively opening up for non-stationary behaviors.

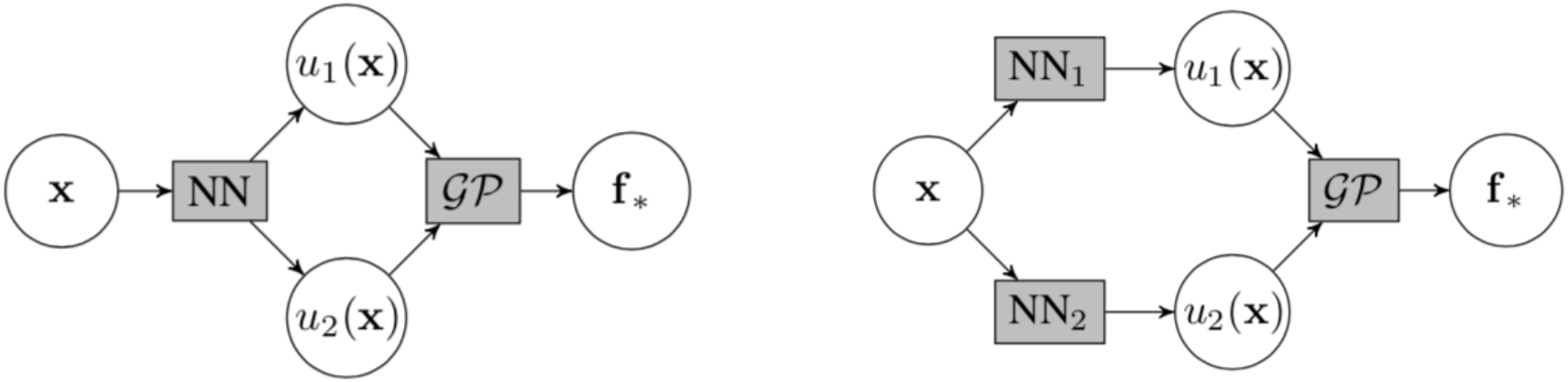
Question: Which mapping should we use?

Let's try a deep neural network...

Roberto Calandra, Jan Peters, Carl E. Rasmussen, and Marc P. Deisenroth. **Manifold Gaussian processes for regression**. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2016.

Andrew G.Wilson, Zhiting Hu, Ruslan R. Salakhutdinov, and Eric P. Xing. **Deep kernel learning**. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

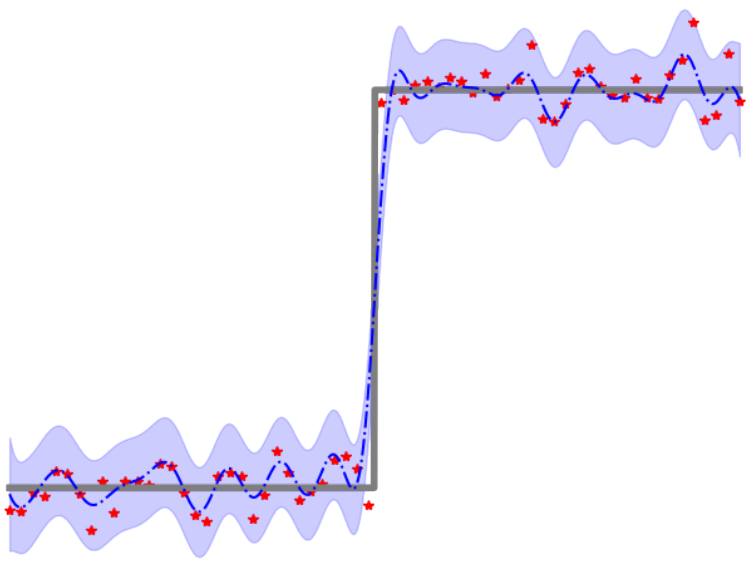
One useful way of combining deep learning with GPs



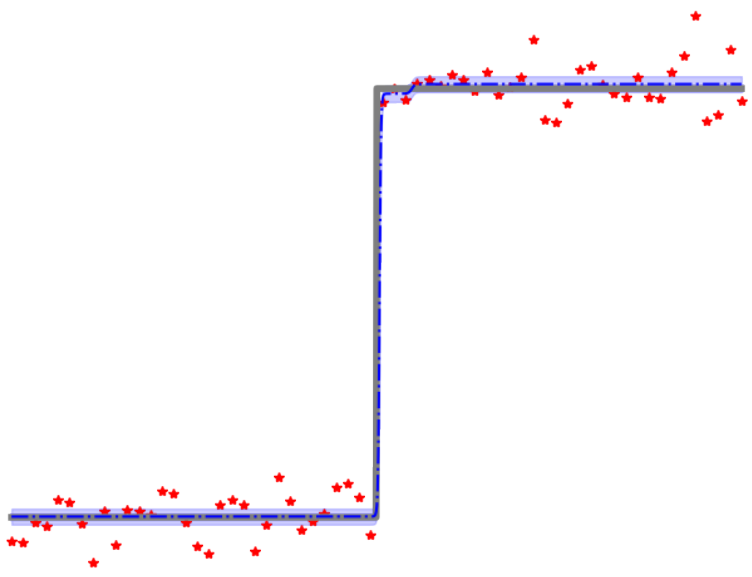
Intuition: The neural network does not have to learn the complete function $f(\mathbf{x})$, but only identify its discontinuities while for the remaining part the model can rely upon the regression capabilities of the GP.

Ex. – illustrating the idea

$$k(x, x') = \sigma_f^2 e^{-\frac{1}{2l^2}(x-x')^2}$$

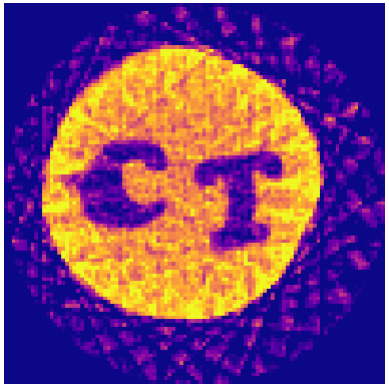


$$k(x, x') = \sigma_f^2 e^{-\frac{1}{2l^2}(u(x)-u(x'))^2}$$

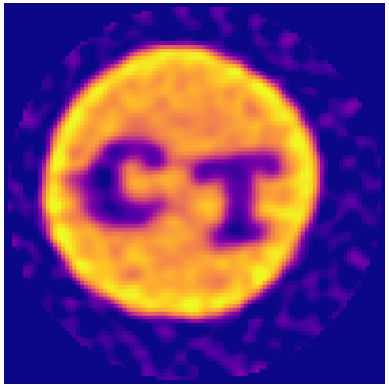


Using the idea together with integral measurements

Ground truth



FBP



GP

GP + DL

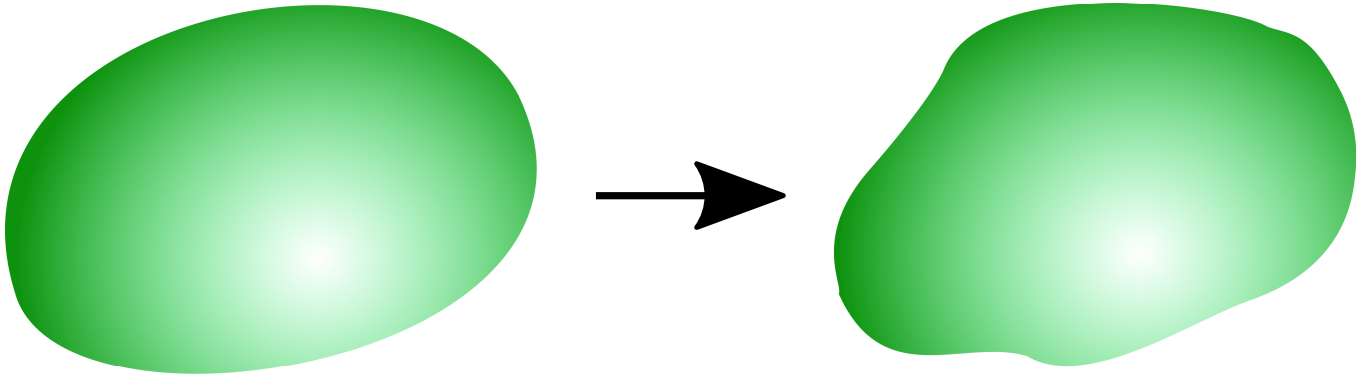
GP + DL: Deep learning to use the input mapping together with our taylored GP prior encoding our understanding of the underlying physics.

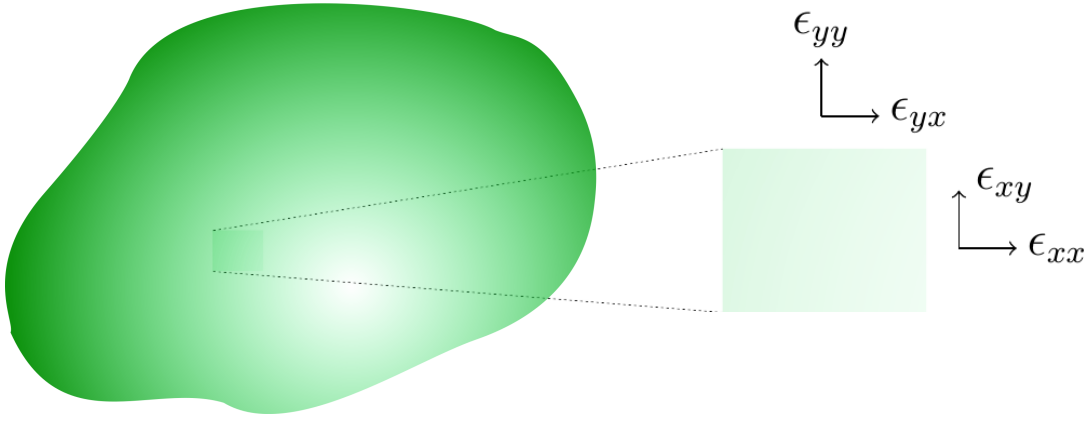
Recall our vision: Create flexible model building blocks containing the basic knowledge we have about the phenomenon we are studying.

Strain field reconstruction – background

Tomographic reconstruction: Recover the internal structure of an object from irradiation experiments.

Deformed object



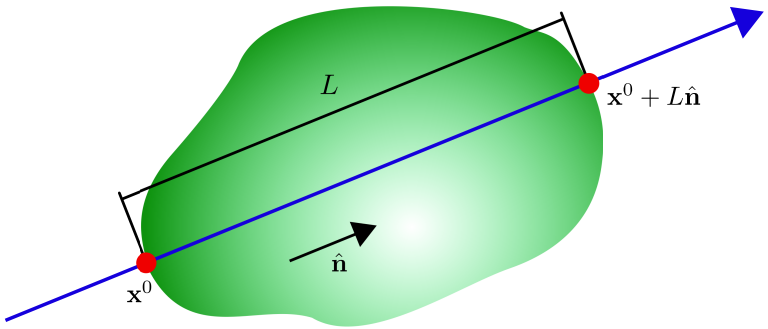


Reconstruct the **strain tensor**

$$\boldsymbol{\epsilon}(\mathbf{x}) = \begin{bmatrix} \epsilon_{xx}(\mathbf{x}) & \epsilon_{xy}(\mathbf{x}) \\ \epsilon_{xy}(\mathbf{x}) & \epsilon_{yy}(\mathbf{x}) \end{bmatrix}$$

Strain field reconstruction – measurement model

Neutron beams are generated at a source, transmitted through the sample (along $\hat{\mathbf{n}}$) and recorded at a detector.



Measurement model (vectorised form):

$$y = \frac{1}{L}\int_0^L \boldsymbol{N}^{\mathsf{T}} \boldsymbol{f}(\mathbf{x}^0 + s\hat{\boldsymbol{n}})ds + \varepsilon$$

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \epsilon_{xx}(\mathbf{x}) \\ \epsilon_{xy}(\mathbf{x}) \\ \epsilon_{yy}(\mathbf{x}) \end{bmatrix}, \qquad \mathbf{N} = \begin{bmatrix} n_x^2 \\ 2n_x n_y \\ n_y^2 \end{bmatrix}$$

Strain field reconstruction – covariance model

Put a GP on the strain field $\mathbf{f}(\mathbf{x})$

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(\mathbf{x}, \mathbf{x}'))$$

Since $\mathbf{f}(\mathbf{x})$ is multivariate, the covariance function is a **matrix**

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} k_{11}(\mathbf{x}, \mathbf{x}') & k_{12}(\mathbf{x}, \mathbf{x}') & k_{13}(\mathbf{x}, \mathbf{x}') \\ k_{21}(\mathbf{x}, \mathbf{x}') & k_{22}(\mathbf{x}, \mathbf{x}') & k_{23}(\mathbf{x}, \mathbf{x}') \\ k_{31}(\mathbf{x}, \mathbf{x}') & k_{32}(\mathbf{x}, \mathbf{x}') & k_{33}(\mathbf{x}, \mathbf{x}') \end{bmatrix}$$

How should we select $\mathbf{K}(\mathbf{x}, \mathbf{x}')$?

There are certain physical constraints that it needs to fulfill.

Multivariate GP – constraint incorporation

Assume linear constraints

$$\mathscr{F}_{\mathbf{x}}\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Let $\mathbf{f}(\mathbf{x}) = \mathscr{G}_{\mathbf{x}}\mathbf{g}(\mathbf{x})$

$$\mathbf{f}(\mathbf{x}) = \mathscr{G}_{\mathbf{x}}\mathbf{g}(\mathbf{x}) \sim \mathcal{GP}\left(\mathscr{G}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{g}(\mathbf{x})},\ \mathscr{G}_{\mathbf{x}}\mathbf{K}_{\mathbf{g}(\mathbf{x},\mathbf{x}')}\mathscr{G}_{\mathbf{x}'}^{\mathsf{T}}\right)$$

Then

$$\mathscr{F}_{\mathbf{x}}\mathscr{G}_{\mathbf{x}}\mathbf{g}(\mathbf{x}) = \mathbf{0}$$

Arbitrary $\mathbf{g}(\mathbf{x})$

$$\Rightarrow \mathscr{F}_{\mathbf{x}}\mathscr{G}_{\mathbf{x}} = \mathbf{0}$$

Find $\mathscr{G}_{\mathbf{x}}$

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.

Multivariate GP – constraint incorporation

Toy Example

Let

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

and consider the constraint

$$\frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y} = 0 \quad \Leftrightarrow \quad \underbrace{\begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{bmatrix}}_{\mathscr{F}_{\mathbf{x}}} \mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Need $\mathscr{G}_{\mathbf{x}}$ such that $\mathscr{F}_{\mathbf{x}}\mathscr{G}_{\mathbf{x}} = \mathbf{0}$. One option is

$$\mathscr{G}_{\mathbf{x}} = \begin{bmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{bmatrix}$$

since

$$\mathscr{F}_{\mathbf{x}}\mathscr{G}_{\mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{bmatrix} \begin{bmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{bmatrix} = -\frac{\partial^2}{\partial x \partial y} + \frac{\partial^2}{\partial y \partial x} = 0.$$

Strain field reconstruction – constraint incorporation

A physical strain field must satisfy the **equilibrium constraints** (isotropic linear elastic solid materials under plain stress)

$$0 = \frac{\partial f_{xx}(\mathbf{x})}{\partial x} + (1-\nu)\frac{\partial f_{xy}(\mathbf{x})}{\partial y} + \nu\frac{\partial f_{yy}(\mathbf{x})}{\partial x},$$
$$0 = \nu\frac{\partial f_{xx}(\mathbf{x})}{\partial y} + (1-\nu)\frac{\partial f_{xy}(\mathbf{x})}{\partial x} + \frac{\partial f_{yy}(\mathbf{x})}{\partial y}.$$

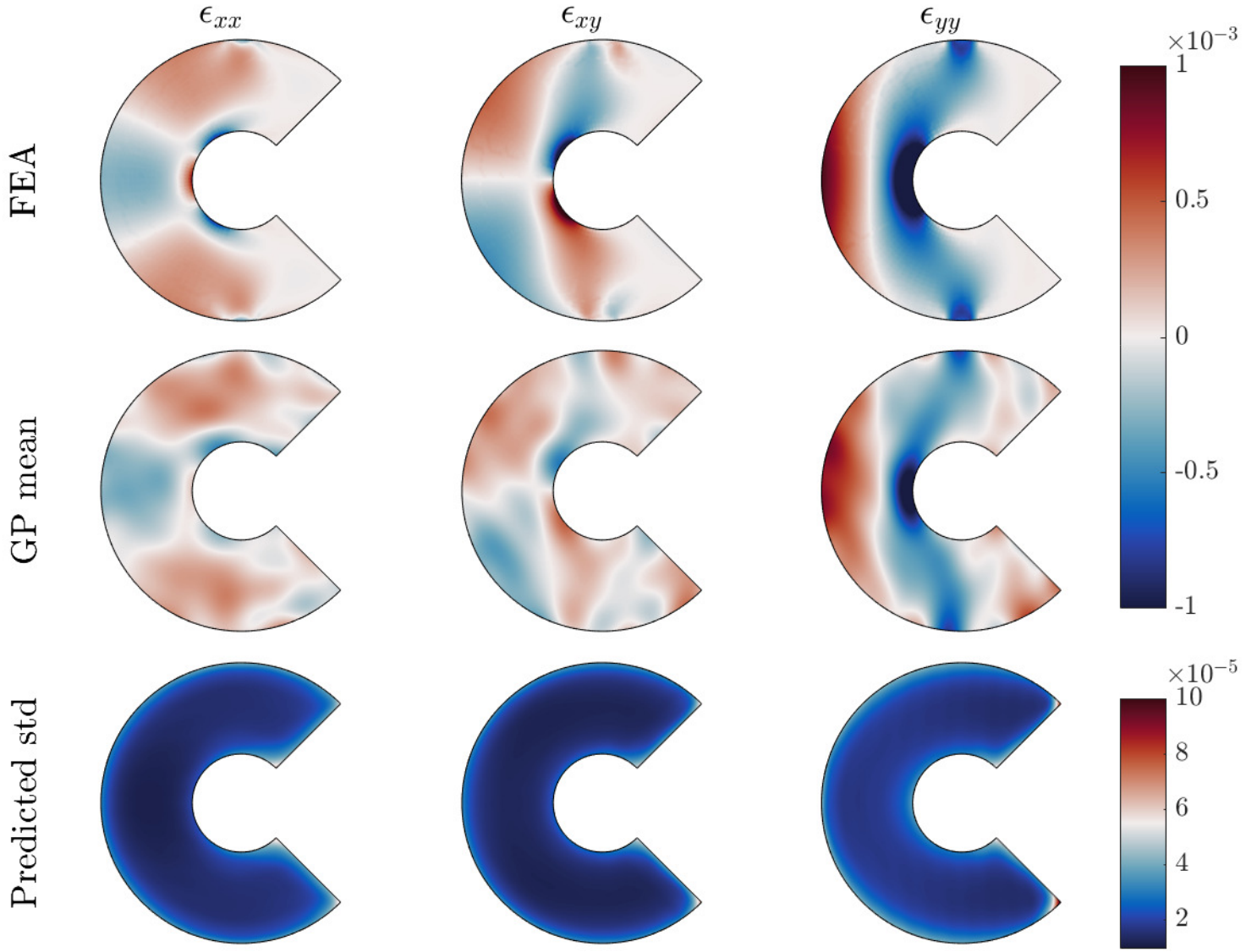
These can be written as

$$\mathbf{0} = \underbrace{\begin{bmatrix} \frac{\partial}{\partial x} & (1-\nu)\frac{\partial}{\partial y} & \nu\frac{\partial}{\partial x} \\ \nu\frac{\partial}{\partial y} & (1-\nu)\frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{bmatrix}}_{\mathscr{F}_{\mathbf{x}}} \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{c}_1^{\mathsf{T}} \\ \boldsymbol{c}_2^{\mathsf{T}} \end{bmatrix} \mathbf{f}(\mathbf{x})$$

We have constructed a Gaussian process that is **guaranteed to obey linear operator constraints** by shaping the covariance function

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

Strain field reconstruction – experimental results



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

Conclusion

The **combined use** of data-driven flexible models and existing knowledge can be quite rewarding.

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

We introduced one flexible model class: Gaussian process (GP)

Hinted at how to embed basic knowledge from physics into the GP.

Uncertainty is a key concept!

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