

# Domain Knowledge in Data Analysis: A Geometrical Method for Low-Dimensional Representations of Simulations

Jochen Garcke



Fraunhofer Centre for Machine Learning

partial support by BMBF Big Data / ML Initiative and BMWi



Barbara Fuchs, Rodrigo Iza-Teran, Sebastian Mayer, Mandar Pathare, Nikhil Prabakaran,  
Sebastian Schmitz, Daniela Steffes-Lai

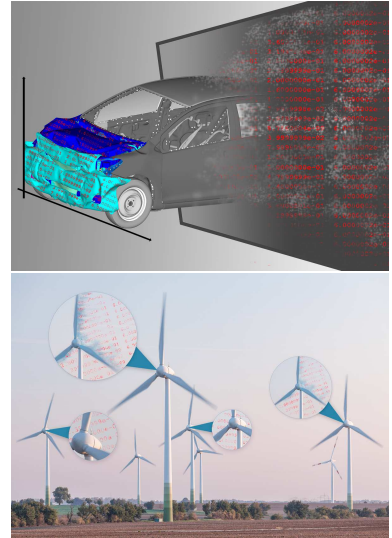
Science at Extreme Scales WS II: HPC and Data Science for Scientific Discovery

# Fraunhofer Gesellschaft at a Glance

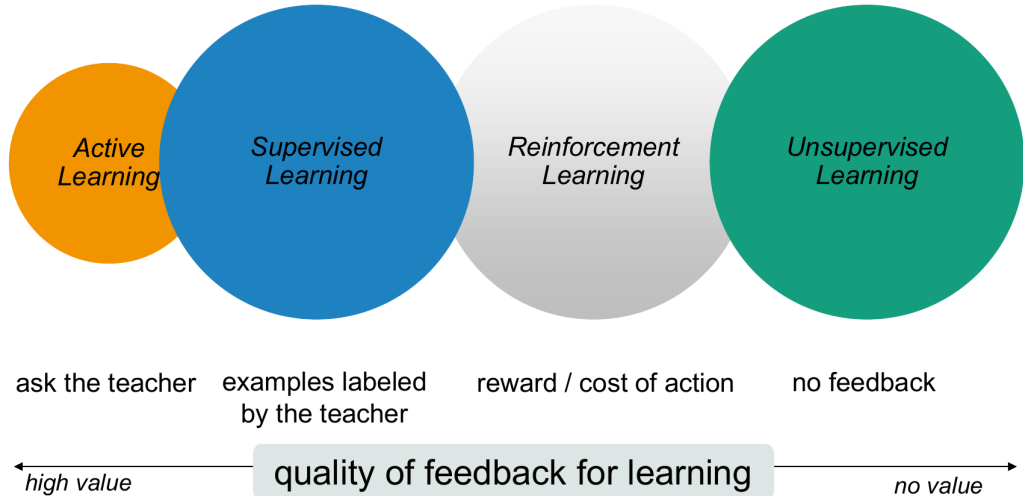
- application-oriented research for immediate benefit to the economy and to the benefit of society
- 72 institutes and research units
- 25,000 staff
- largest organization for applied research in Europe
- 2.3 € billion annual research budget totaling. Of this sum, more than 2.0 € billion is generated through contract research
  - roughly two thirds of this sum is generated through contract research on behalf of industry and publicly funded research projects
  - roughly one third is contributed by the German federal and Countries governments in the form of base funding
- several Fraunhofer subsidiaries and centers worldwide

# Machine Learning at Fraunhofer SCAI

- We generate machine learning methods
  - develop new machine learning approaches
  - integrate domain knowledge into machine learning algorithms
  - improve scalability of machine learning methods
- We adapt machine learning for and use it in applications
  - industrial (virtual) product development
  - data-driven energy management for networks
  - predictive maintenance
  - design of innovative materials
  - interpretation of patient data
  - knowledge graph / taxonomy by text mining of medical publications
  - anomaly detection in (telecommunication) logs
- We provide training and coaching for machine learning

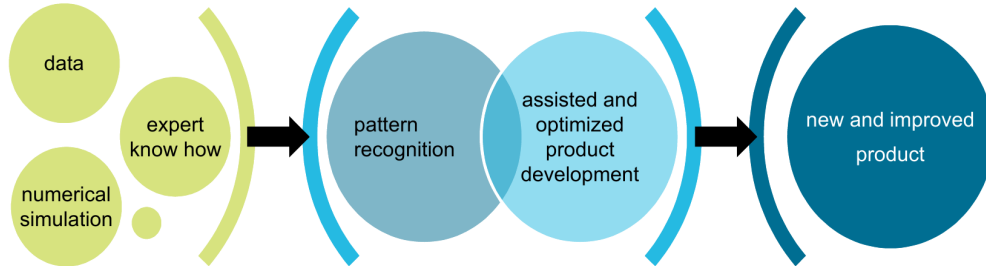


# Machine Learning Domains



# Machine Learning in (Virtual) Product Development

- machine learning tools allow analysis of complex data arising
    - from highly detailed numerical simulations during (virtual) product development
    - from sensors / sensor network / control data
  - use ML to **simplify** data analysis in R&D process and **assist** development engineer
  - for complex physical data aim for a **structured integration** of domain know-how into ML
- machine learning**

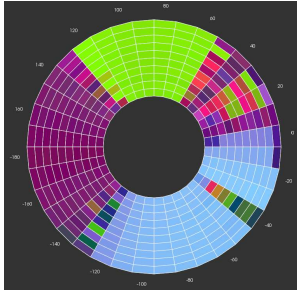


- **The lack of information cannot be remedied by any mathematical trickery.** Lanczos 1961

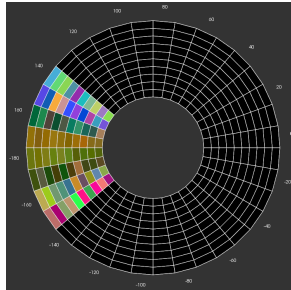
# Machine Learning in Design of Wind Energy Plant Controller

- time series from simulations of WEP arise in their design, fine-tuning of installations or upgrades
- meta data encompass environmental and operational conditions (specified by certification bodies), as well instantiations of wind turbine's individual components
- due to complexity and volume of this raw data, automated post-processing is used, i.e. to identify anomalies or the overrun of thresholds (research project with GE Lab Garching)

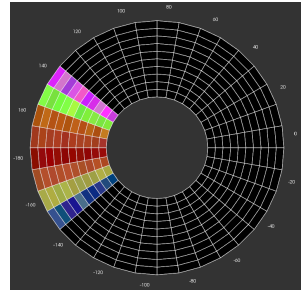
ED:



ED:



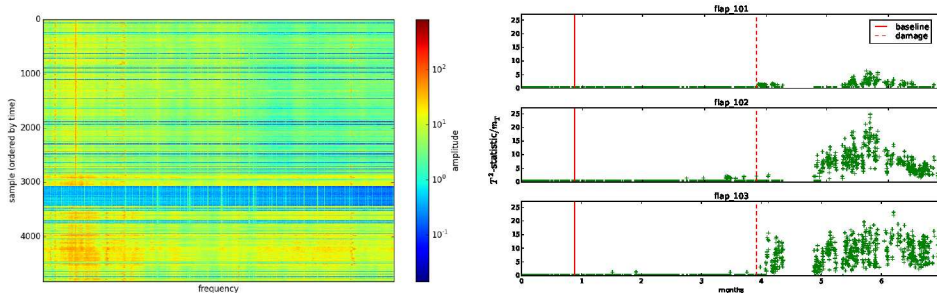
DTW:



- use numerical simulations to generate data for controller design for rare / catastrophic cases

# Condition Monitoring for Wind Energy Plants

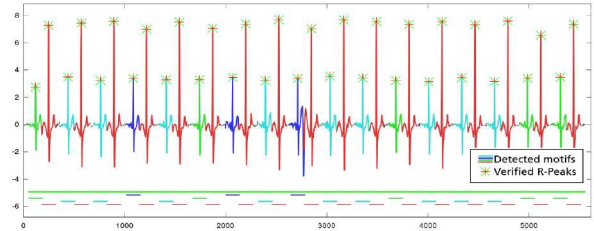
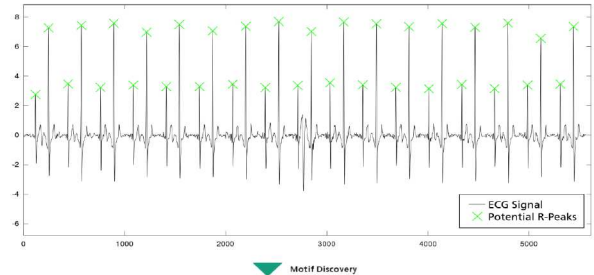
- analysis of sensor data out of condition monitoring system for wind energy plant
- hourly frequency measurements from vibration sensors on blades
- history over several years from Weidmüller Monitoring Systems, originally for ice detection
- no stable learning phase due to strong effect of wind
- investigated case: early detection of damages in rotor blades, based on historical data of defects



- in MADESI: use numerical simulations with sensor data to study aging of blades / gear boxes

# Anomaly Detection in Time-Dependent Patient Data

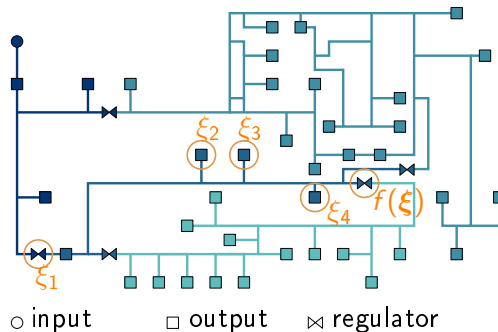
- joint work with Universitätsklinikum Bonn
- analysis of ECG time series
- challenges
  - strong noise and external influences
  - ECG shape depends on the patient
  - expensive labelling by hand
  - anomaly detection without a stable learning phase
- focus on detection of QRS complex
- our approach developed with clinicians
  - motif discovery for regular events in time series
  - visualisation of detected motifs categorizes data
- aim is to reduce time needed for observing live data and reduce workload for medical staff





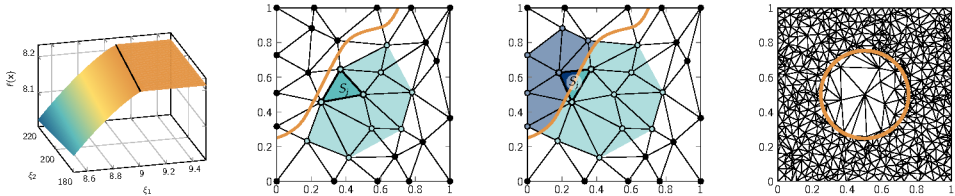
# Uncertainty Quantification for Gas Networks - Setup

- scenario analysis necessary to operate gas network safely and reliably
  - How much gas does each customer withdraw? What happens for demand peaks ?
  - influence of temperature around pipe (only roughly / uncertainly known for a time frame)
  - roughness of pipe (cannot be measured in a working pipe)
- interested in expectation or cummulative distribution function of pressure in pipe

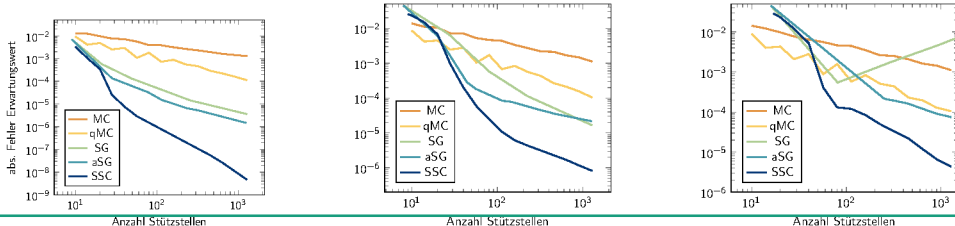


# Uncertainty Quantification for Gas Networks - Results

- kinks in surrogate surface can arise due to pressure regulation  $\rightarrow$  piecewise smooth function
- we can check and exploit whether regulator is active or not (after each simulation run)



- our adaptive sampling approach yields better results than other common methods (2d to 4d)



# CRISP-DM

## An Infrastructure for Business Analytics

- Cross Industry Standard Process for Data Mining
- Industry consortium, v1: 1996
- breaks process of data mining into six phases
- leading methodology used by industry data miners



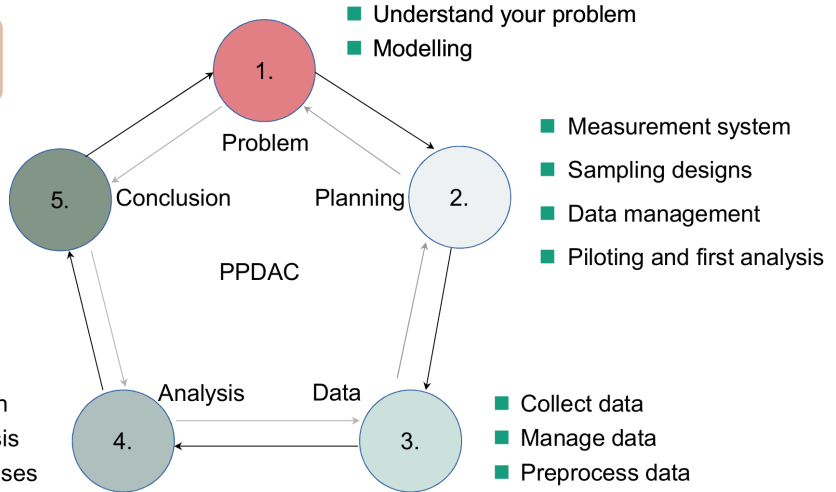
[en.wikipedia.org/wiki/Cross-industry\\_standard\\_process\\_for\\_data\\_mining](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining)

# The statistical method (MacKay & Oldford 2002)

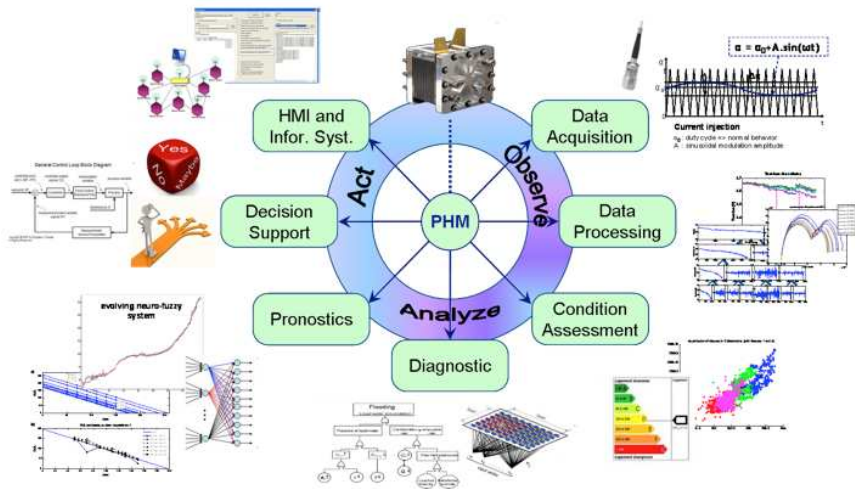
**Retrospection:** that's how  
Statistics is used in science

- Interpretation
- Conclusions
- New ideas
- Communication

- Data exploration
- Targeted analysis
- Testing hypotheses

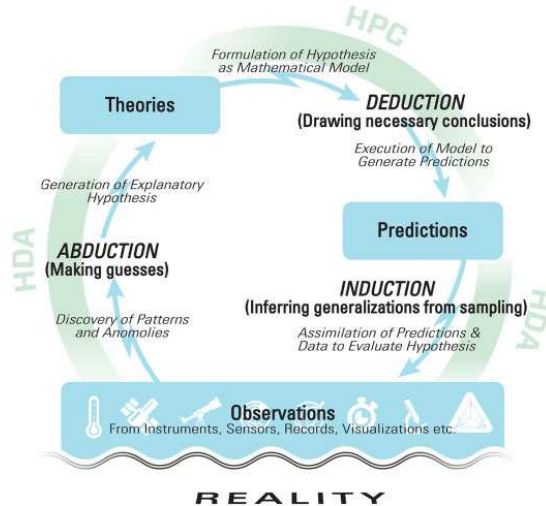


# System Monitoring / Predictive Health Monitoring



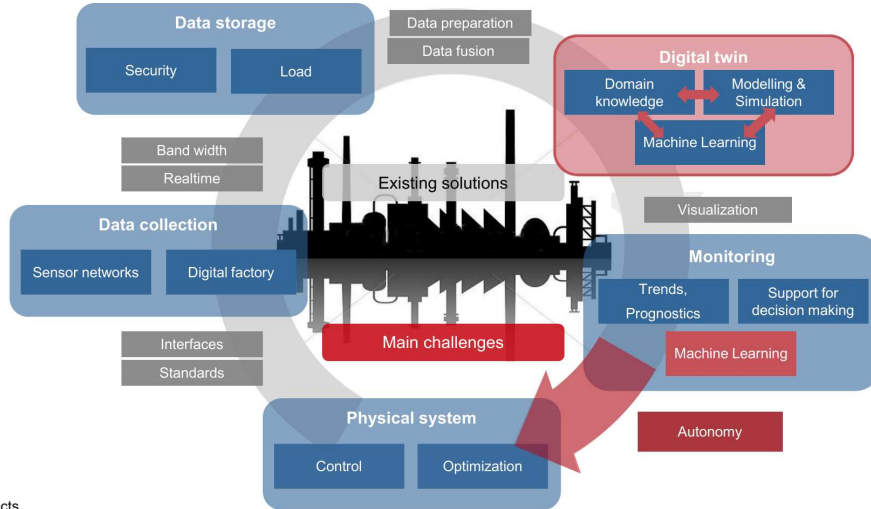
[www.fclab.fr/wp-content/uploads/2012/08/phm\\_axis\\_fclab.png](http://www.fclab.fr/wp-content/uploads/2012/08/phm_axis_fclab.png)

# Inference Cycle for the Process of Scientific Inquiry



Big data and extreme-scale computing DOI:10.1177/1094342018778123

# Industry 4.0: Full-fledged CPS need fusion of data and simulation



# The Clash of the Cultures ?

Industry 4.0 implies  
marrying of two worlds, which work and function under very different rules

**(German)**  
**machinery- and plant engineering**

high precision

errors are expensive

(defective goods; high material costs)

deployment involves many planning stages  
(this is definitely not scrum)

processes do not easily just scale

**(American)**  
**software development**

errors are at most annoying for the customer:  
deploy patches

errors do not immediately cost money:  
lower requirements for precision

agile development

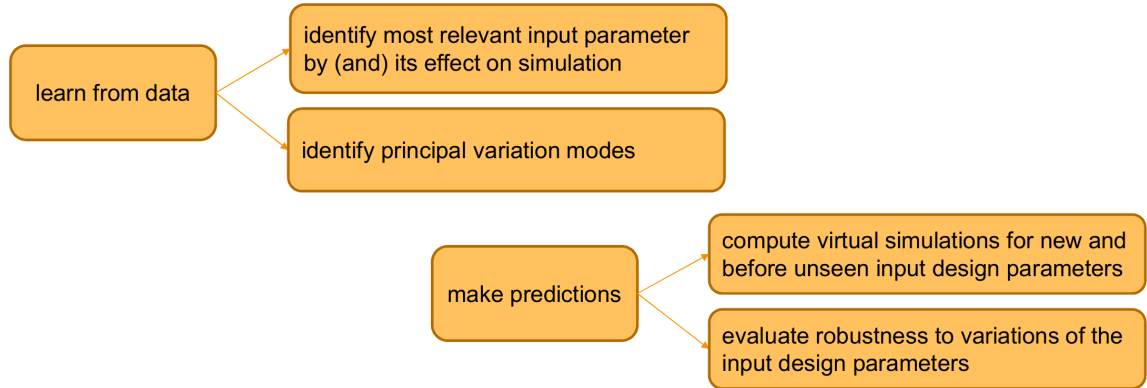
scalability is in parts almost for free



# Machine Learning for CAE ?

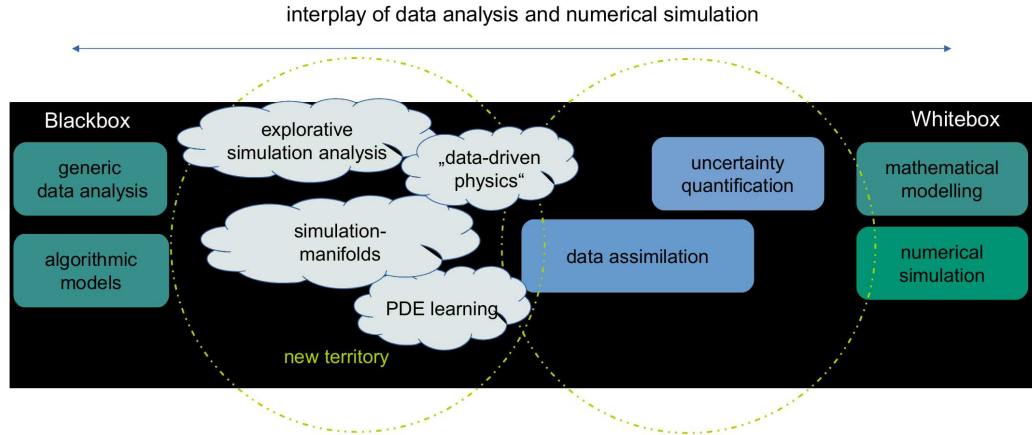
from a machine learning definition:

“..study/construct algorithms that can learn from and make predictions on data”



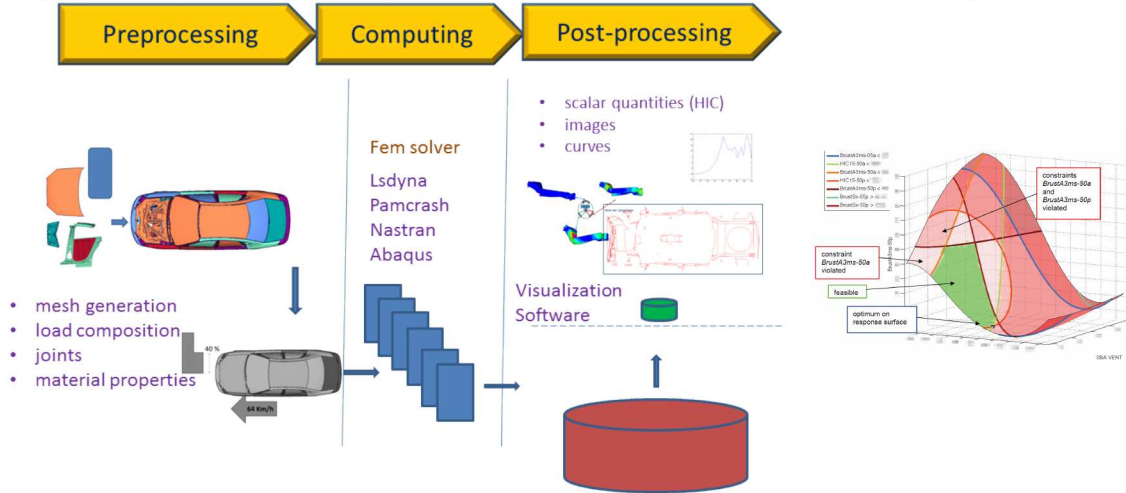
# Greybox Machine Learning: Methodological brother of digital twins

- What is the role of data analysis in modelling and simulation?
- How do expert knowledge and machine learning intertwine?



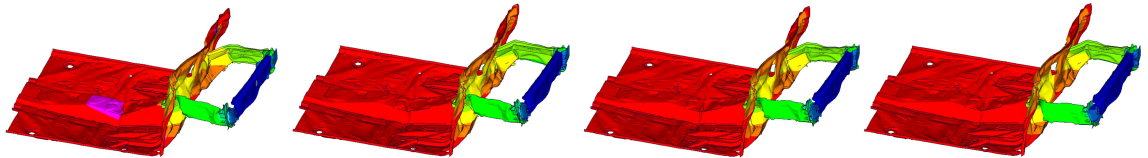
# Virtual Product Development with CAE in Automotive

- highly developed & regulated (load cases from: EuroNCAP, FMVSS, ECE-R,...)



# Analysis of Data from Crash Simulations

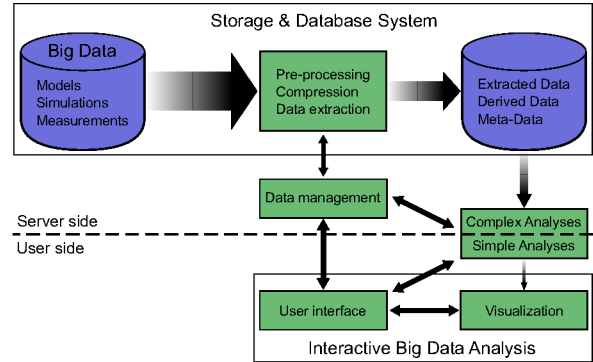
- design quantities such as thickness, geometry, material properties (or its modelling)
- per R&D-step can arise (couple of) hundred simulation runs
- response surface / data analysis for scalars (e.g. head injury criterion, firewall intrusion)
- for detailed analysis data needs to be investigated interactively
- visualisation for single 3D simulation, but no tools to compare **geometric deformations**



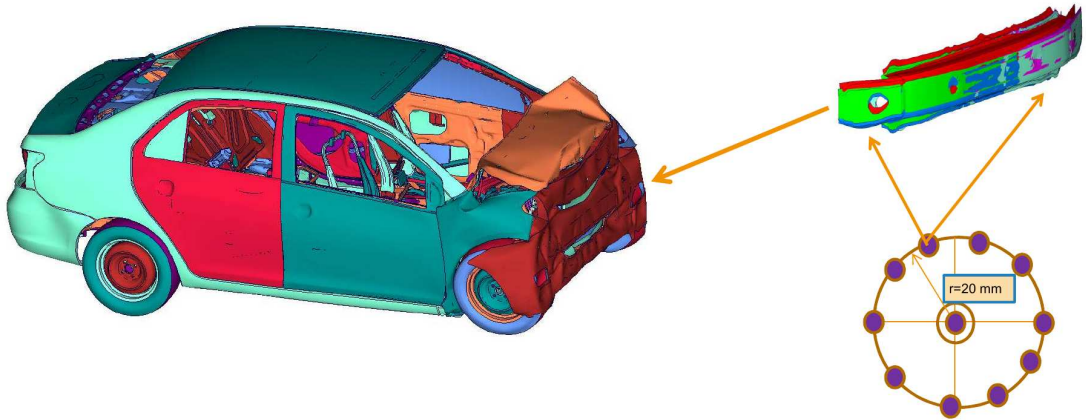
- our aim: automatic **organisation** of several (full) simulation results
- goal: find intrinsic dimension  $s$  of  $d$ -dimensional simulation vectors,  $s \ll d$

# Handling of Bundles of Numerical Simulation Data

- simulation data is **bulky data**, therefore
  - not stored in database but in special file formats (often vendor specific)
  - results are organised “database-like” using simulation data management (SDM)-systems
  - store meta data, derived data, etc. with simulation data
- for analysis data needs to be **easily accessible** (storage, transfer, visualisation)
- goal: employ data storage server
  - compress simulation data (e.g. FEMZIP)
  - compute mainly at data, not at client
  - exploit HPC capabilities of server
- future goal for product development
  - integrate sensor and measurement data
  - align real and simulation data

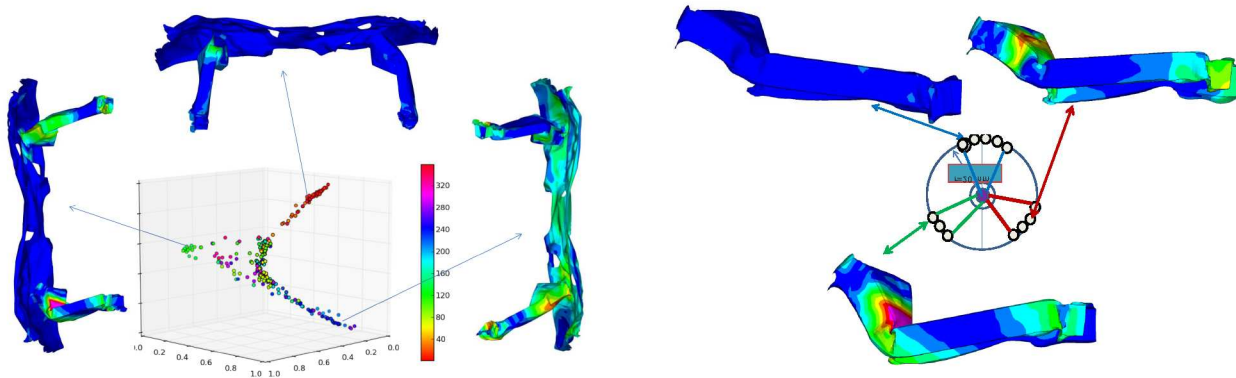


# Study on Position of Bumper for Toyota Yaris



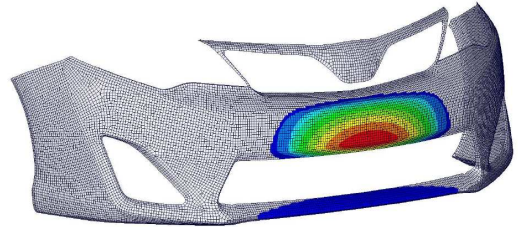
- bumper attachment positions parametrised by angle, varied on a small circle
- perform simulation for ca. 200 parameters

# Low Dimensional Organisation of Simulations by Diffusion Maps

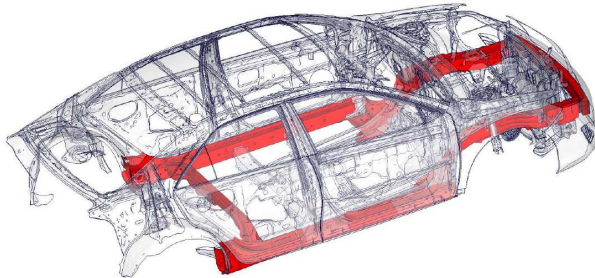


- investigate array of absolute deformations on firewall and structural beams
- each point represents a simulation result, color coding according to nodal distance to reference
- the identified three dominant modes correspond to three angle regions
- engineer selects wanted / unwanted deformation behavior in embedding → classification labels

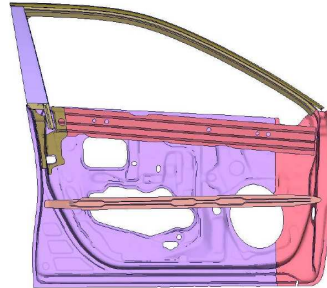
- FEM model comparison based on semantic segmentation
  - changes to rigid body elements
  - changes to spotwelds
  - duplicate parts (translated / rotated parts)
  - new / missing parts and elements



identification of geometry and mesh changes



detection of material and thickness changes



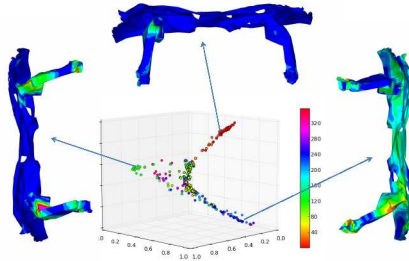
multi-parts detection



# Dimensionality Reduction / Simulation Space

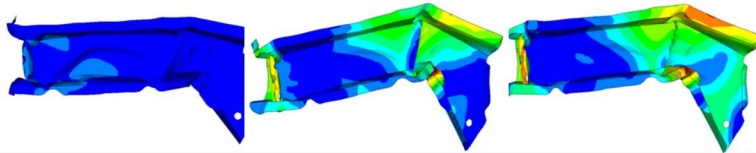
- simulations are high dimensional objects

Manifold Learning



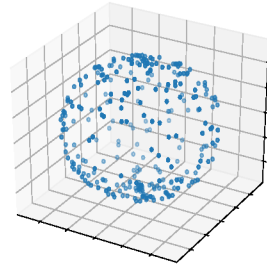
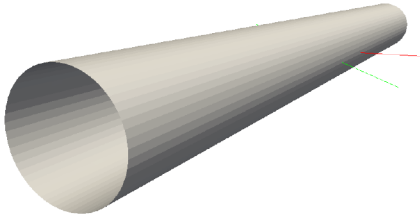
- simulations are transformed from reference

Orbit Space



# Mathematical Motivation: Orbit Space

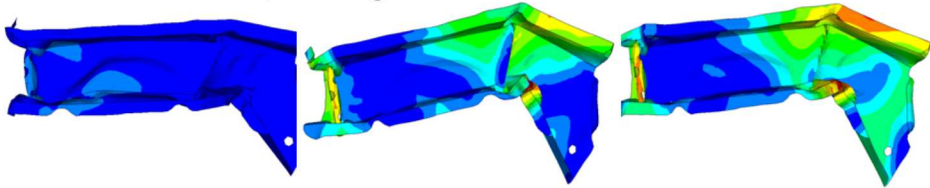
- assume simulations are obtained by transformation from reference simulation  $f_0$ 
  - $f = \gamma \cdot f_0$ ,  $\gamma \in G$  with  $f, f_0 \in \mathcal{M}$
- parametrize simulations according to such transformations
  - $\mathcal{M}$  space of all simulations objects
  - $\mathcal{M}/G$  space of simulations modulo a transformation group
  - $G \cdot f := \{(\gamma, f) \mid \gamma \in G\}$  is the orbit



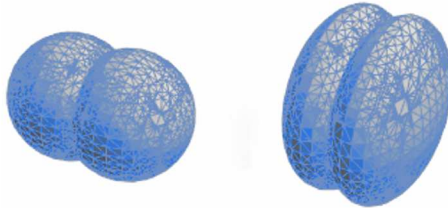
- exploit  $G$  to understand the space of simulations objects  $\mathcal{M}$
- study objects invariant under group of transformations  $G$

# Symmetry - Structure Preservation in Transformation of Objects

- isometric invariant  $\rightarrow$  distance preserving



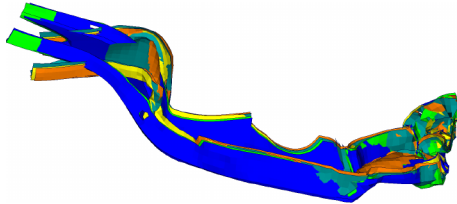
- affine invariant  $\rightarrow$  collinearity preserving



- conformal invariant  $\rightarrow$  angle preserving

# Invariance for Simulation Bundles

- although no closed form available, principle can be used  
invariance
- different simulations have different surface deformations
- variability is in many cases preserving the distance on the surface, i.e. an isometry  
look for distance preserving operator



- Laplace-Beltrami operator is distance preserving on a mesh
- data-driven Fokker-Planck operator is invariant to nonlinear transformation, due to its construction from observed data

# Discrete Laplace Beltrami Operator

- For a mesh  $K$  which is an  $(\epsilon, \eta)$  approx. of a surface  $\mathcal{S}$ , (Belkin, Sun, and Wang, 2008) defined for any vertex  $w$  the mesh Laplace-Beltrami operator (with  $d(p, w)$  the graph distance)

$$L_K^h f(w) = \frac{1}{4\pi h^2} \sum_{t \in K} \frac{\text{Area}(t)}{\#t} \sum_{p \in V(t)} e^{-\frac{d(p,w)^2}{4h}} (f(p) - f(w)),$$

(Belkin, Sun, and Wang, 2008) showed for  $f \in C^2(\mathcal{S})$  and suitable  $h(\epsilon, \eta)$

$$\lim_{\epsilon, \eta \rightarrow 0} \sup_{K(\epsilon, \eta)} \left\| L_K^{h(\epsilon, \eta)} f - \Delta_{\mathcal{S}} f|_K \right\|_{\infty} = 0,$$

where the supremum is taken over all  $(\epsilon, \eta)$ -approximations of  $\mathcal{S}$ .

- recent works give point wise estimates between graph Laplacians and continuum operators or their spectral convergence
- for eigenvectors and eigenprojections (so far) only consistency results are known

# Invariant Operators: Isometric Invariance

## Theorem (Iza-Teran, G., 2018)

*Let the set of meshes  $\bar{K} = \{K^i\}_{i=1}^m$  contain the approximations of the surfaces  $S^i$  and let there be transformations  $\varphi|_K, i = 1, \dots, m$  between the meshes which are  $\varepsilon$ -isometric. Assume that the variance for the  $\varepsilon$ -isometric transformations follow a Gaussian distribution. Then, the approximation of the Laplace-Beltrami operator  $L_K^h$ , constructed using graph distances for one mesh  $K$ , differs only by a scaling factor from the ones for the deformed meshes  $i = 1, \dots, m$ .*

## sketch.

Geodesic distance stays the same after an isometric transformation, calculating the Laplace-Beltrami operator based on it will lead to the same result. The geodesic distance is now approximated by the graph distance, where we assume a small error perturbation of magnitude  $\varepsilon$  which follows a Gaussian distribution. Now, we can use a result from random matrix theory, which states that for the Gaussian kernel, such a Graph Laplacian matrix disturbed by noise can be considered a rescaled version of the original one. □

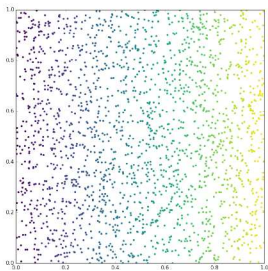
# Spectral Decomposition of an Operator

- on a manifold  $(S, g)$  define eigenvalue problem  $-\Delta_S \psi = \lambda \psi$
- operator is p.s.d., eigenvalues  $\lambda_k$ ,  $k \geq 0$  are real positive and isolated with finite multiplicity
- use corresponding discrete operator and its discrete eigenfunctions  $\{\psi_i\}_i^N$
- strongly related to the use of Laplace-Beltrami operator in **shape analysis / shape spaces**
- spectral decomposition in operator eigenbasis gives for function  $f$  on mesh

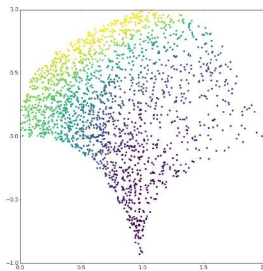
$$f = \sum_{i=1}^N \alpha^i \psi_i, \alpha^i = \langle f, \psi_i \rangle$$

- distance of coefficients  $\alpha_1^i, \alpha_2^i$  gives good distance measure for corresponding simulations  $f^1, f^2$

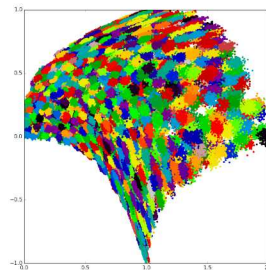
# Fokker-Planck Operator Constructed From Data



non-observable data



observed data



simulation burst

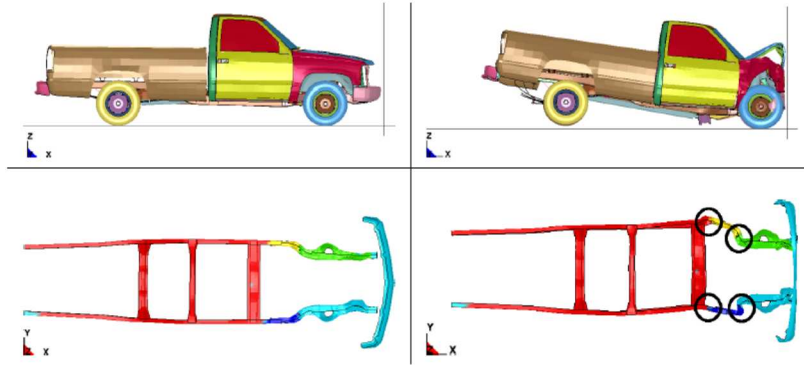
- estimate local covariance matrices  $C$  from observed cloud, use  $C = J_\varphi J_\varphi^T$
- (Singer and Coifman, 2008): approximate original distance for  $p$ 's, up to  $O(\|\eta - \eta'\|_{R^d}^2)$

$$d(p, p')^2 := 2(\eta - \eta')^T \left[ J_\varphi J_\varphi^T(\eta) + J_\varphi J_\varphi^T(\eta') \right]^{-1} (\eta - \eta')$$

- use **data-driven distances as graph weights**, build Fokker-Planck operator  $L = W_{rs} - I$
- discrete operator is invariant to nonlinear transformation  $\varphi$

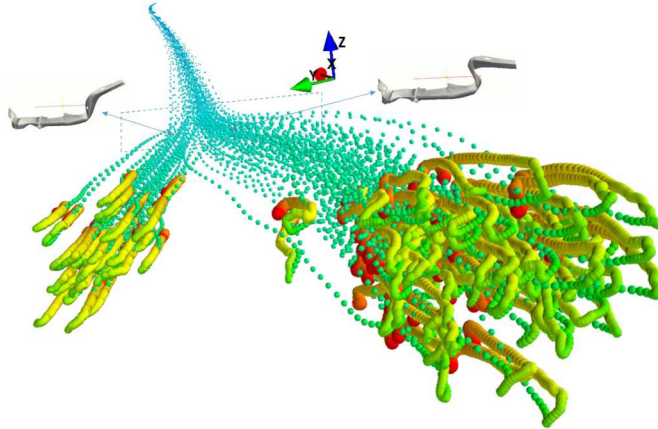


# Analysis of Numerical Simulations of a Car Crash



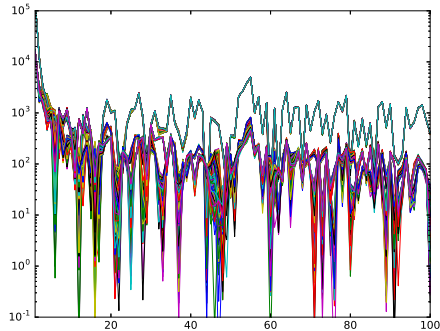
- using numerical simulations where thickness values of 9 parts are varied up to 30%
- chose relevant structural part (and time step) for analysis
- analyse the different deformations of the simulations results

# Visualization of All Time Steps in LB-decomposition

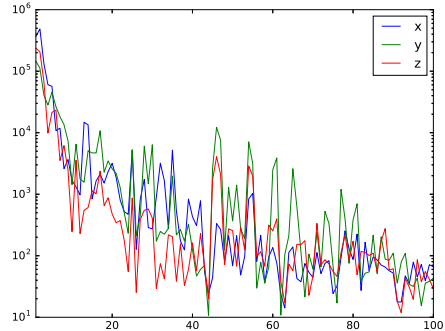


- for each mesh point its  $x$ ,  $y$ , and  $z$  coordinates in simulation  $i$  gives  $f_x^i$ ,  $f_y^i$ , and  $f_z^i$
- use first spectral coefficient for each  $f_x^i$ ,  $f_y^i$ ,  $f_z^i$  at each time

# Fokker-Planck operator



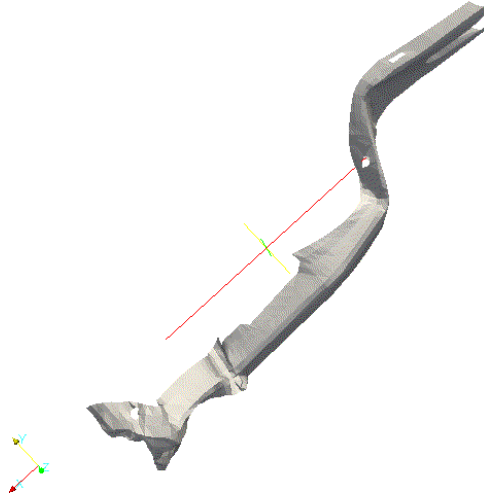
(a) magnitude of coefficients



(b) variance of coefficients

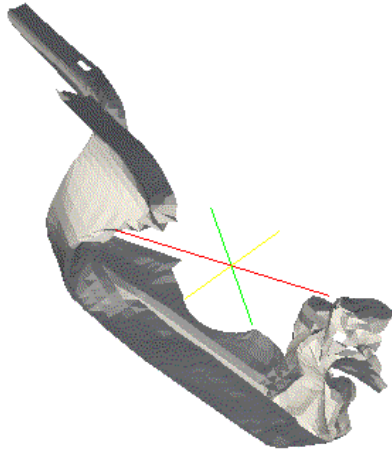
- for each mesh point its  $x$ ,  $y$ , and  $z$  coordinates in simulation  $i$  gives  $f_x^i$ ,  $f_y^i$ , and  $f_z^i$
- spectral decomposition computed for a selected time step
- observation: 1st mode reflects translation, 2nd mode reflects rotation

## Mode 3 - Global Deformation



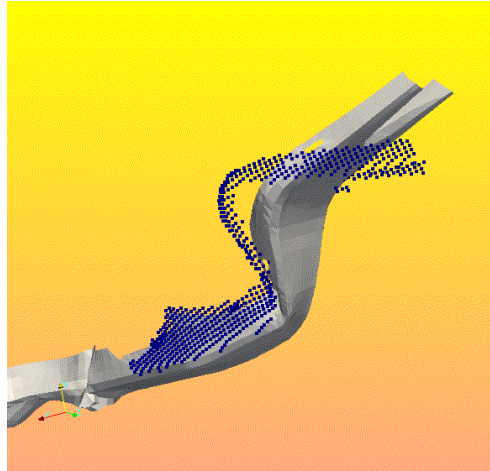
- fix all coefficients but the third one

# Mode 4 - Local Deformation



- fix all coefficients but the fourth one

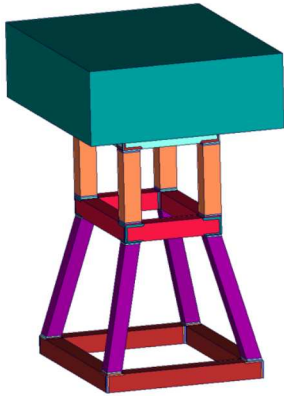
# Morph in Lower Dimensional Representation to Match a Point Cloud



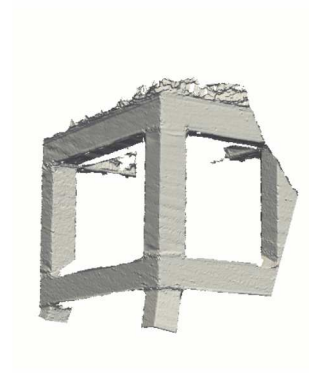
- vary / optimize over several spectral coefficients

# High Speed 3D-Point Cloud Measurements

results from joint project with Fraunhofer IOF and EMI



“Hand”-Build Test Structure



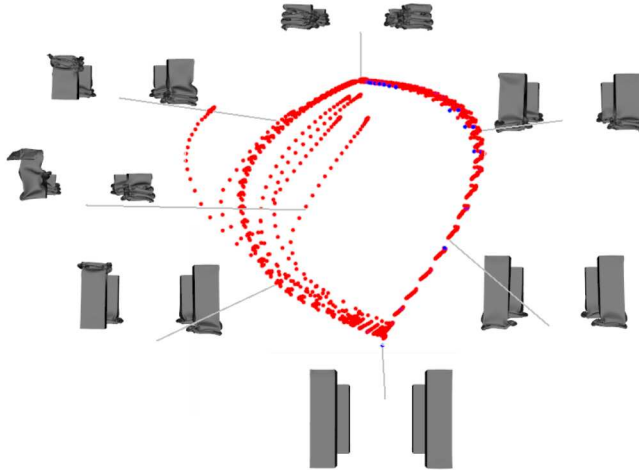
3D-Video of Crash

# Matching of 3D-Point Data and Simulation

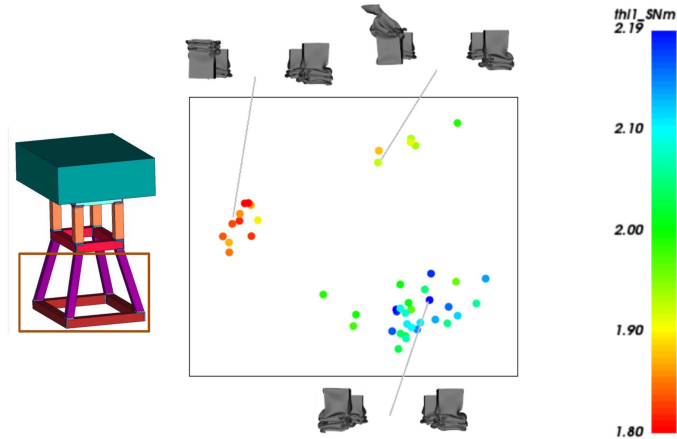




# Path of Experiment Data in Simulation Space



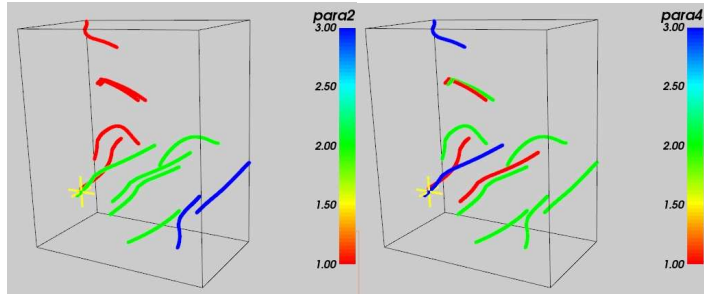
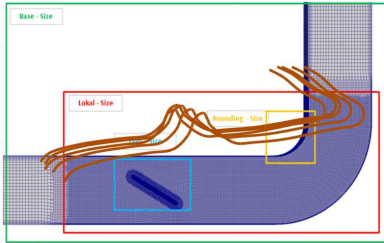
# Analysis of Effect on von Mises-Stress



find matching simulation: obtain information about von Mises-Stress on non-observed structure

# Analysis of CFD Simulation / Aeroacoustic Study on HVAC Channel

- study to find “good” discretization parameters, i.e. good enough, but not too fine
- treat simulation field over middle plane of cavity as surface
- only preliminary results so far inside EU-project Fortissimo 2 with CFD-Schuck









# Conclusion

- integration of domain knowledge and/or assumptions into overall data analysis for complex engineering data essential and possible
- introduce decomposition of invariant operator for analysis of bundles of numerical simulations
- gives **joint basis for bundles of simulations**, e.g. all can be visualised in time together
- allows “virtual” simulations by interpolation in lower dimensional representation
- **use suitable distance measures** on mesh to build domain assumptions into data representation
- could use some more theory, e.g. in regard to approximation of eigenfunctions
- further explore connection to shape spaces in computer graphics
- is there a connection to representation theory ?
- generalization from surface to full 3D data ?
- preliminary results using an invariant basis in RBM-context

**HDA2019: 8th Workshop on High-Dimensional Approximation**

save the date: **9 – 13 September 2019** @ETH Zurich

-  Aguilera, A. et al. (2016). “Advancing a Gateway Infrastructure for Wind Turbine Data Analysis”. In: *Journal of Grid Computing* 14.4, pp. 499–514. DOI: [10.1007/s10723-016-9376-9](https://doi.org/10.1007/s10723-016-9376-9).
-  Belkin, M., J. Sun, and Y. Wang (2008). “Discrete Laplace Operator on Meshed Surfaces”. In: *Proceedings of the Symposium on Computational Geometry*. SoCG '08. College Park, MD, USA: ACM, pp. 278–287.
-  Bohn, B. et al. (2013). “Analysis of Car Crash Simulation Data with Nonlinear Machine Learning Methods”. In: *Procedia Computer Science, Proceedings of the ICCS 2013, Barcelona*. Vol. 18, pp. 621–630. DOI: [10.1016/j.procs.2013.05.226](https://doi.org/10.1016/j.procs.2013.05.226).
-  Fuchs, B. and J. Garcke (2018). *Simplex Stochastic Collocation for Piecewise Smooth Functions with Kinks*. almost submitted.
-  Garcke, J. and R. Iza-Teran (2017). “Machine Learning Approaches for Data from Car Crashes and Numerical Car Crash Simulations”. In: *NAFEMS 2017, Stockholm*.
-  Garcke, J., R. Iza-Teran, et al. (2017). “Dimensionality Reduction for the Analysis of Time Series Data from Wind Turbines”. In: *Scientific Computing and Algorithms in Industrial Simulations: Projects and Products of Fraunhofer SCAI*. Springer, pp. 317–339. DOI: [10.1007/978-3-319-62458-7\\_16](https://doi.org/10.1007/978-3-319-62458-7_16).



Garcke, J., M. Pathare, and N. Prabakaran (2017). “ModelCompare”. In: *Scientific Computing and Algorithms in Industrial Simulations: Projects and Products of Fraunhofer SCAI*. Springer, pp. 199–205. DOI: 10.1007/978-3-319-62458-7\_10.



Iza-Teran, R. and J. Garcke (2018). *A Geometrical Method for Low-Dimensional Representations of Simulations*. revision submitted.



Singer, A. and R. R. Coifman (2008). “Non-linear independent component analysis with diffusion maps”. In: *Applied and Computational Harmonic Analysis* 25.2, pp. 226–239. DOI: 10.1016/j.acha.2007.11.001.