16th AIP Open Seminar

10 March 2021

Identifying Memorable Experiences of Learning Machines

Dharmesh Tailor

Paul Chang

Arno Solin

Siddharth Swaroop

Emtiyaz Khan

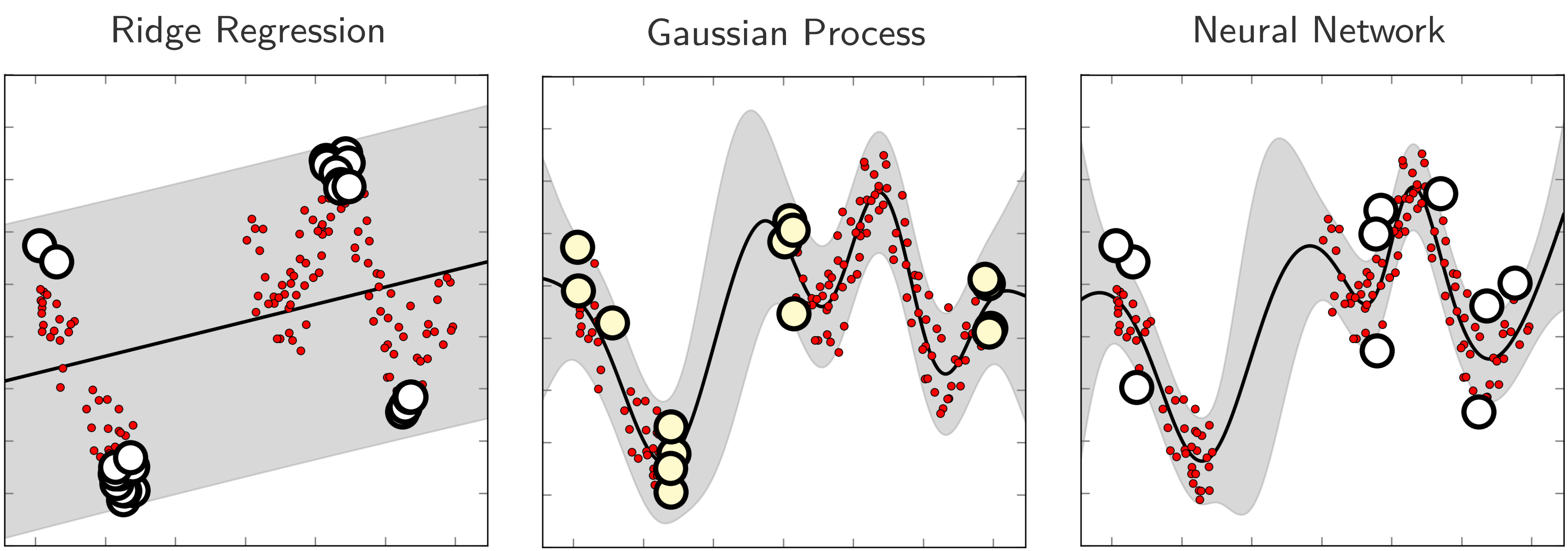
Aalto University

UNIVERSITY OF CAMBRIDGE

RIKEN AIP

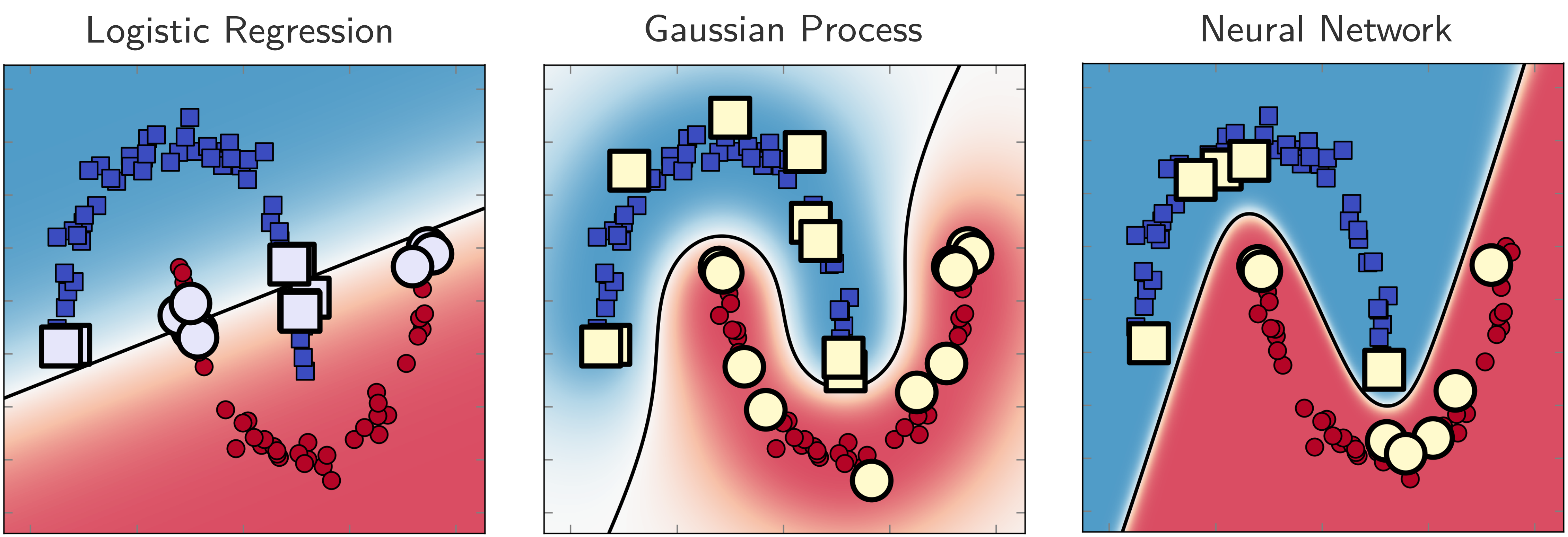
Memorable experiences

- Humans have the ability to identify memorable experiences
- The memorable experiences of a variety of machine-learning models can be identified with a **single Bayesian principle**



Memorable experiences

- Humans have the ability to identify memorable experiences
- The memorable experiences of a variety of machine-learning models can be identified with a **single Bayesian principle**



Related Work

- Influential datapoints
  - Regression diagnostics [1]
  - Influence function [2]
- Sparse Gaussian Processes
  - Variational learning of inducing inputs [3]
  - Subset-of-data approaches [4,5]
- Support vectors [6]
- Coresets [7]

Memorable experiences unify and generalise these concepts by using a single Bayesian principle.

[1] Cook, R. D., & Weisberg, S. *Residuals and influence in regression*. New York: Chapman and Hall. 1982.
[2] Koh, P. W., & Liang, P. *Understanding black-box predictions via influence functions*. ICML, 2017.
[3] Titsias, M. *Variational learning of inducing variables in sparse Gaussian processes.* AISTATS, 2009.
[4] Lawrence, N., et. al. *Fast sparse Gaussian process methods: The informative vector machine*. NeurIPS, 2003.
[5] Burt, D. R., et. al. *Convergence of Sparse Variational Inference in Gaussian Processes Regression.* JMLR, 2020.
[6] Vapnik, V. N. *An overview of statistical learning theory*. IEEE transactions on neural networks, 1999.
[7] Borsos, Z., et. al. *Coresets via Bilevel Optimization for Continual Learning and Streaming*. NeurIPS, 2020.

Ridge Regression

$$\mathbf{w}_* = \min_{\mathbf{w}} \sum_{i=1}^{N} \underbrace{\frac{1}{2}\left(y_i - \mathbf{x}_i^\top \mathbf{w}\right)^2}_{\ell\left(y_i, \mathbf{x}_i^\top \mathbf{w}\right)} + \frac{1}{2}\left\|\mathbf{w}\right\|^2$$

By Lagrangian duality and a variant of the Representer theorem [1]:

$$\mathbf{w}_* = \mathbf{X}^\top \boldsymbol{\alpha}_*$$

Model view *Data view*

$$\boldsymbol{\alpha}_* = \mathbf{y} - \mathbf{X}\mathbf{w}_*$$ *residual*

$$\alpha_{*,i} = -\nabla_{f_i} \ell\left(y_i, f_i\right)\Big|_{f_i = \mathbf{x}_i^\top \mathbf{w}_*}$$

Ridge leverage scores: [2]

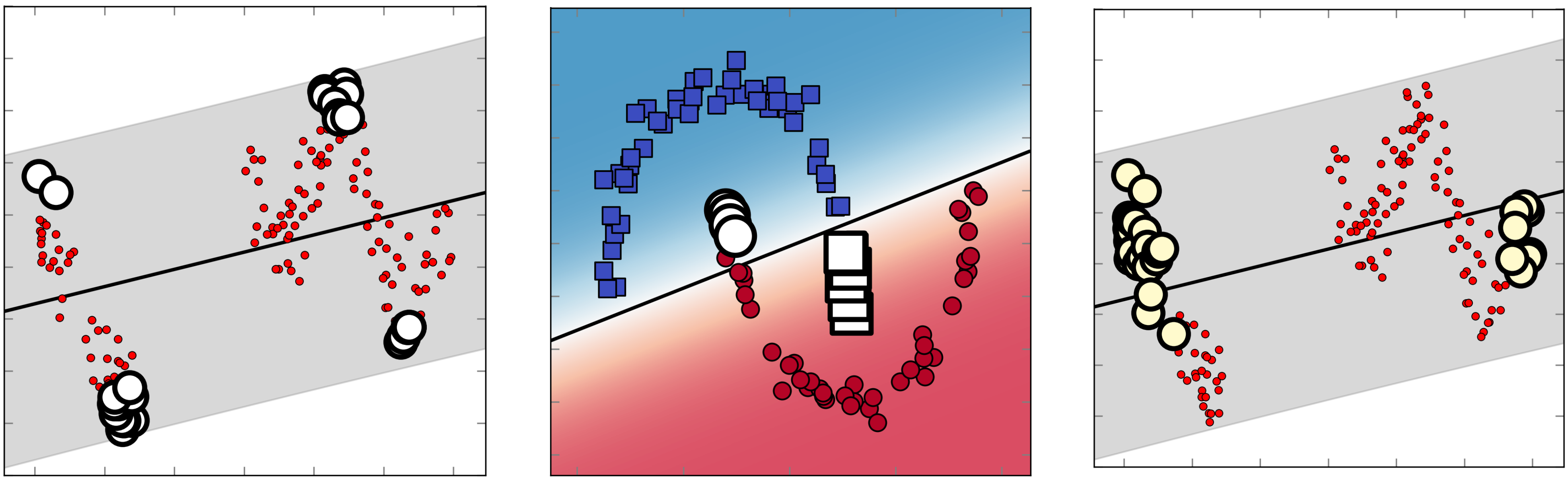
$$\mathbf{f}_{\mathbf{X}} = \mathbf{H}\mathbf{y}$$

$$h_i = \left[\underbrace{\mathbf{X}\mathbf{X}^\top}_{\mathbf{K}}\left(\underbrace{\mathbf{X}\mathbf{X}^\top}_{\mathbf{K}} + \mathbf{I}\right)^{-1}\right]_{ii}$$

[1] Schölkopf, B., et. al. A generalized representer theorem. In International conference on computational learning theory. Springer, 2001.
[2] Alaoui, A. E., & Mahoney, M. W. Fast randomized kernel methods with statistical guarantees. NeurIPS, 2015.

Ridge Regression (and logistic regression)

residual *leverage score*



Gaussian Process

$$\sum_{i=1}^{N} \boxed{\mathbb{E}_{q(f_i)}\left[\log p(y_i \mid f_i)\right]} - \mathrm{D}_{\mathrm{KL}}\left[q(\mathbf{f}) \,\|\, p(\mathbf{f})\right]$$

Gaussian posterior approximation: $q(\mathbf{f}) := \mathrm{N}(\mathbf{f} \mid \mathbf{m}, \mathbf{V})$

Prior: $p(\mathbf{f}) := \mathrm{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K})$

Fixed point of the variational objective:

residual $\mathbf{m}_* = \mathbf{K}\boldsymbol{\alpha}_*$ $\qquad \alpha_{*,i} := \mathbb{E}_{q_*(f_i)} \boxed{\left[-\nabla_{f_i} \ell\left(y_i, f_i\right)\right]}$

$$\mathbf{V}_* = \left[\mathbf{K}^{-1} + \boldsymbol{\Lambda}_*\right]^{-1} \qquad \Lambda_{*,ii} := \mathbb{E}_{q_*(f_i)} \boxed{\left[\nabla^2_{f_i f_i} \ell\left(y_i, f_i\right)\right]}$$

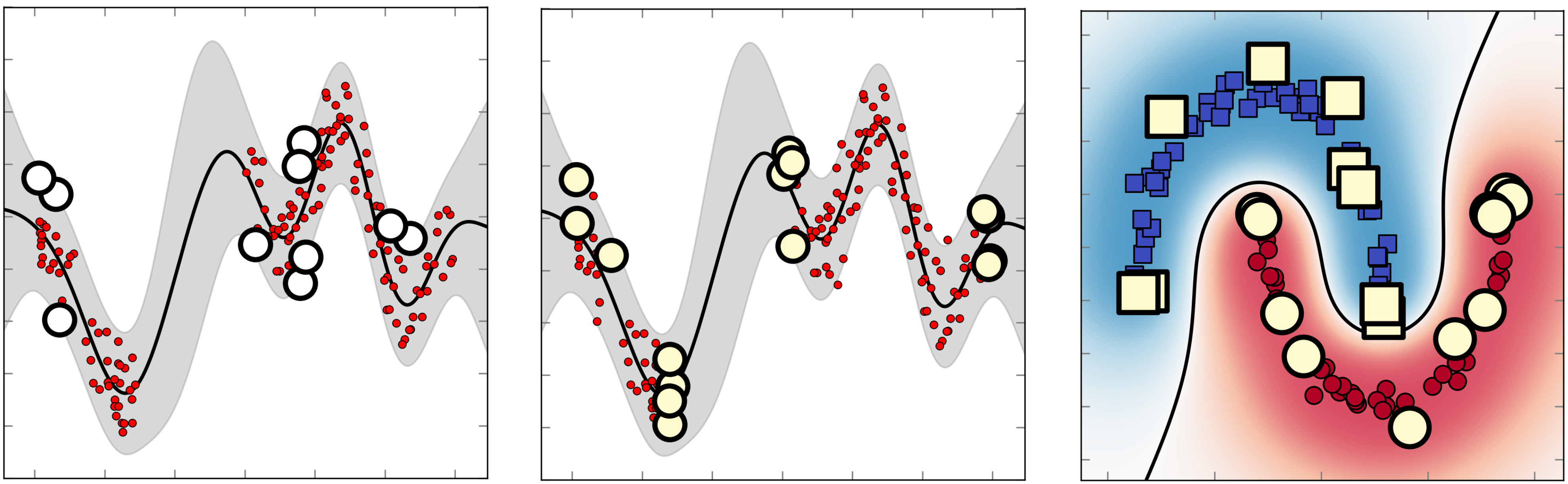
$$\ell\left(y_i, f_i\right) := -\log p(y_i \mid f_i)$$

leverage score $h_i = \left[\mathbf{K}\left(\mathbf{K} + \boldsymbol{\Lambda}_*\right)^{-1}\right]_{ii}$

Khan, M.E., et. al. (2013). **Fast dual variational inference for non-conjugate latent gaussian models.** In International conference on machine learning.

Gaussian Process

residual *leverage score*



Neural Network

$$\mathbb{E}_{q(\mathbf{w})}\left[\sum_{i=1}^{N} \boxed{\ell\left(y_i, f_{\mathbf{w}}(\mathbf{x}_i)\right)} + \frac{1}{2}\|\mathbf{w}\|^2\right] - H[q(\mathbf{w})]$$

Gaussian posterior approximation: $q(\mathbf{w}) := \mathrm{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{V})$

Solution of the Bayesian learning problem:

residual

$$\mathbf{m}_* = \mathbf{J}^\top \boldsymbol{\alpha}_* \qquad \alpha_{*,i} = \boxed{-\nabla_{f_i} \ell\left(y_i, f_i\right)}$$

$$\mathbf{V}_* = \left[\mathbf{J}^\top \boldsymbol{\Lambda}_* \mathbf{J} + \mathbf{I}\right]^{-1} \qquad \Lambda_{*,ii} = \boxed{\nabla^2_{f_i f_i} \ell\left(y_i, f_i\right)}$$

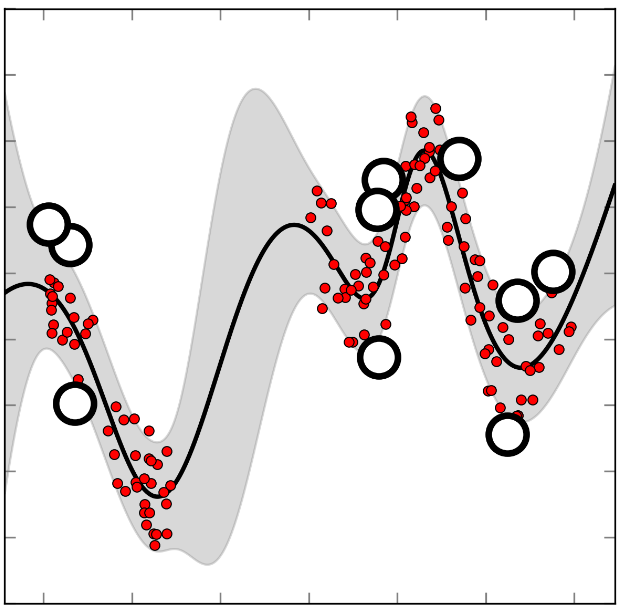
leverage score

$$h_i = \left[\mathbf{J}\mathbf{J}^\top \left(\mathbf{J}\mathbf{J}^\top + \boldsymbol{\Lambda}_*\right)^{-1}\right]_{ii}$$

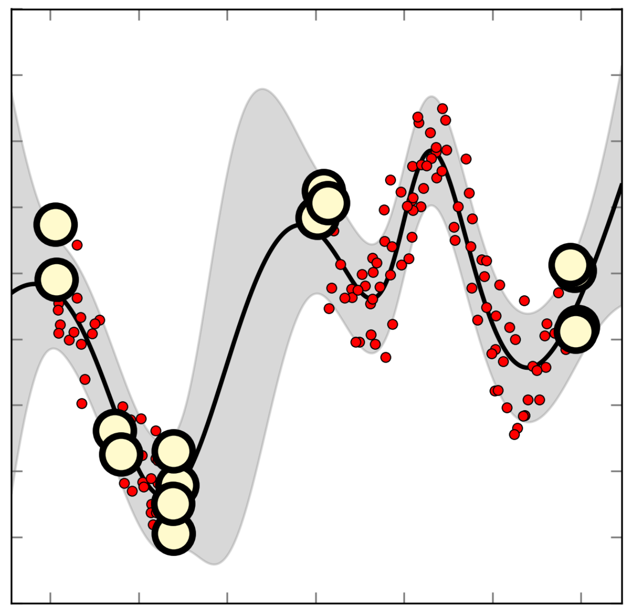
Khan, M. E., et. al. (2019). **Approximate inference turns deep networks into gaussian processes.** In Advances in neural information processing systems.

Neural Network

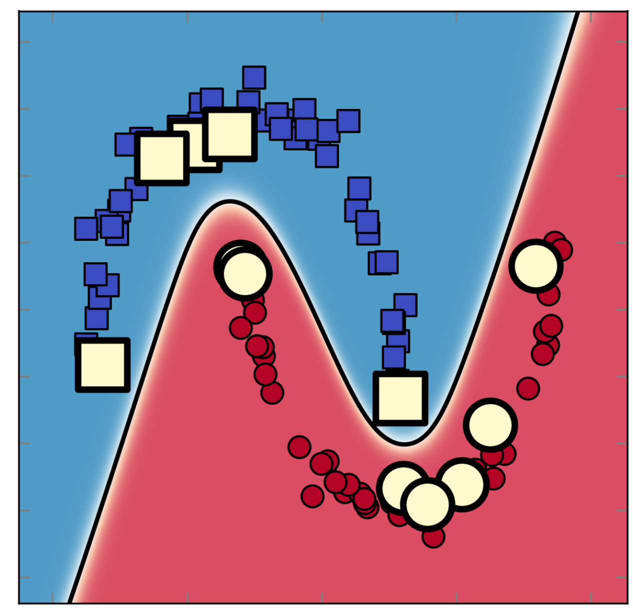
residual



leverage score



lambda



Characterizing memorable experiences

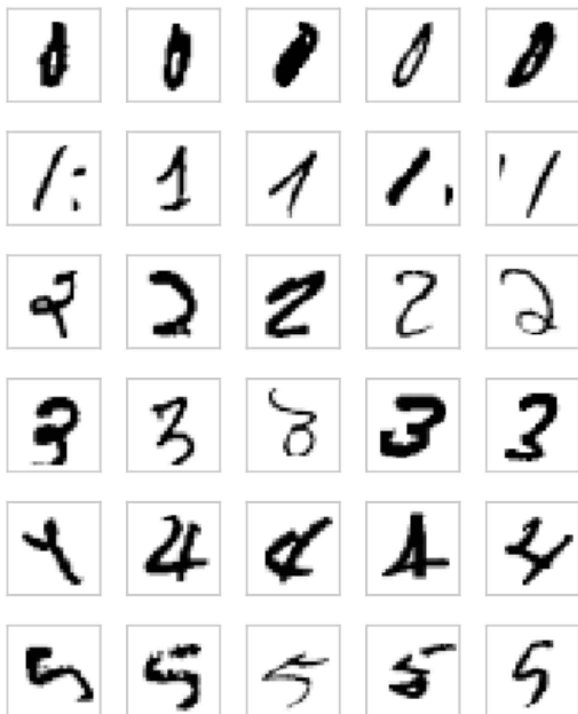
- Choice of criterion depends on application, for example:
  - In lifelong learning scenario (with no task boundaries), examples at boundary of data space may be preferred $\rightarrow$ ***leverage score***
  - Identifying examples for further inspection (e.g. mislabelled) $\rightarrow$ ***residual***

Characterizing memorable experiences

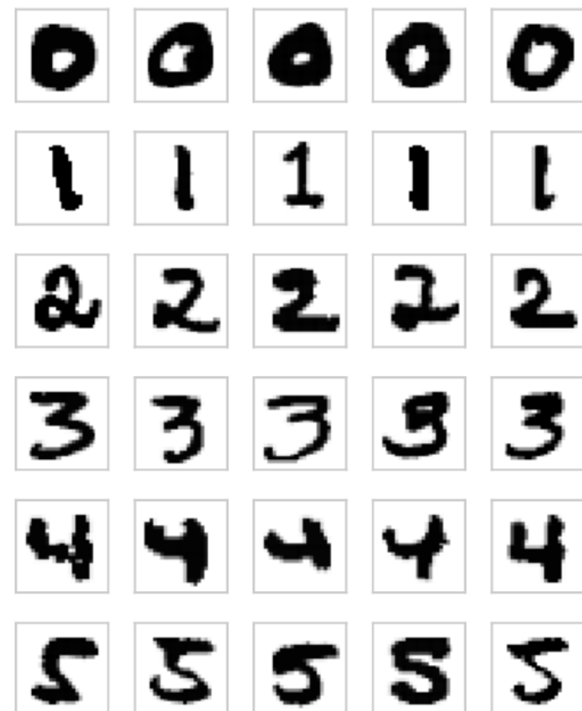
- Continual learning with task boundaries, seek to maintain decision boundary as move to new tasks $\rightarrow \lambda$

MNIST

Most memorable



Least memorable



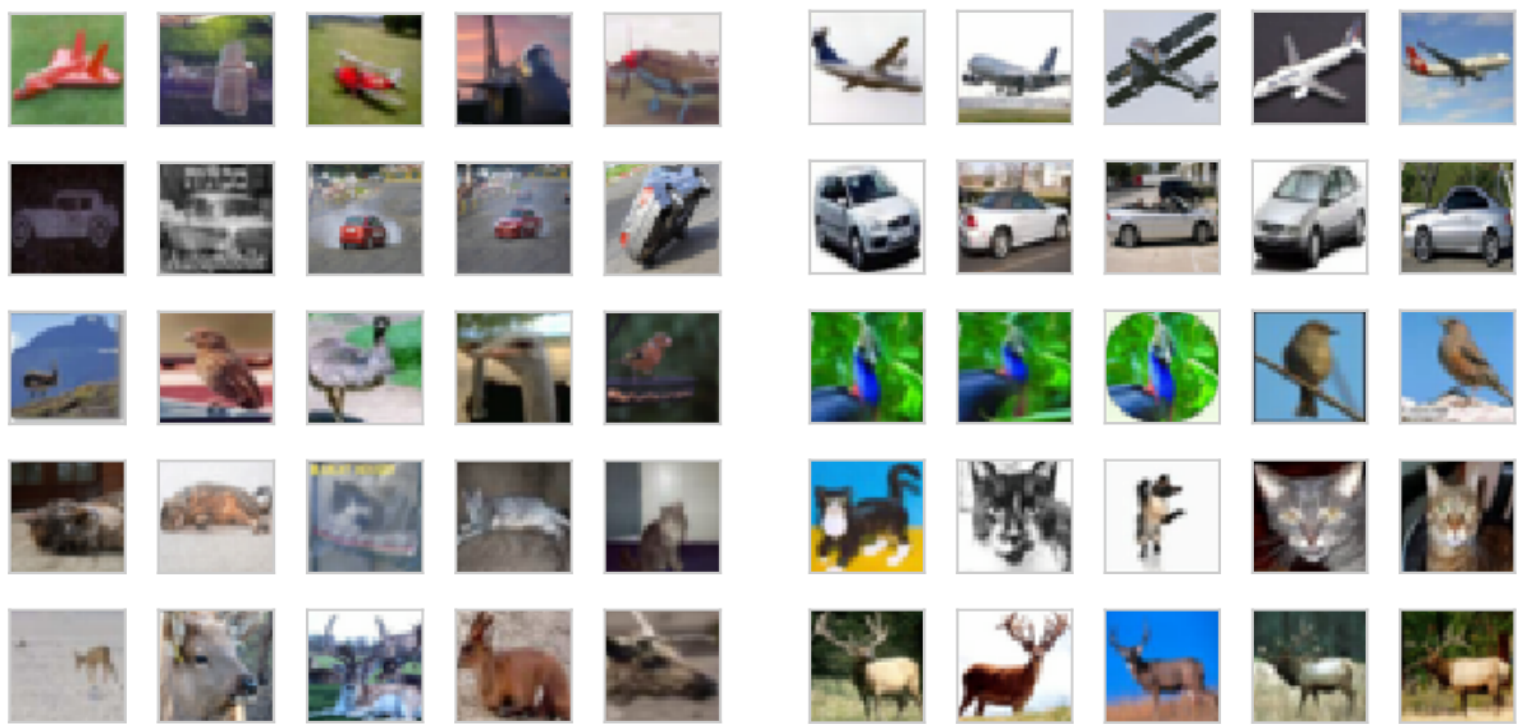
Pan, P., et. al. **Continual deep learning by functional regularisation of memorable past.** NeurIPS, 2020.

Characterizing memorable experiences

- Continual learning with task boundaries, seek to maintain decision boundary as move to new tasks $\rightarrow \lambda$

CIFAR

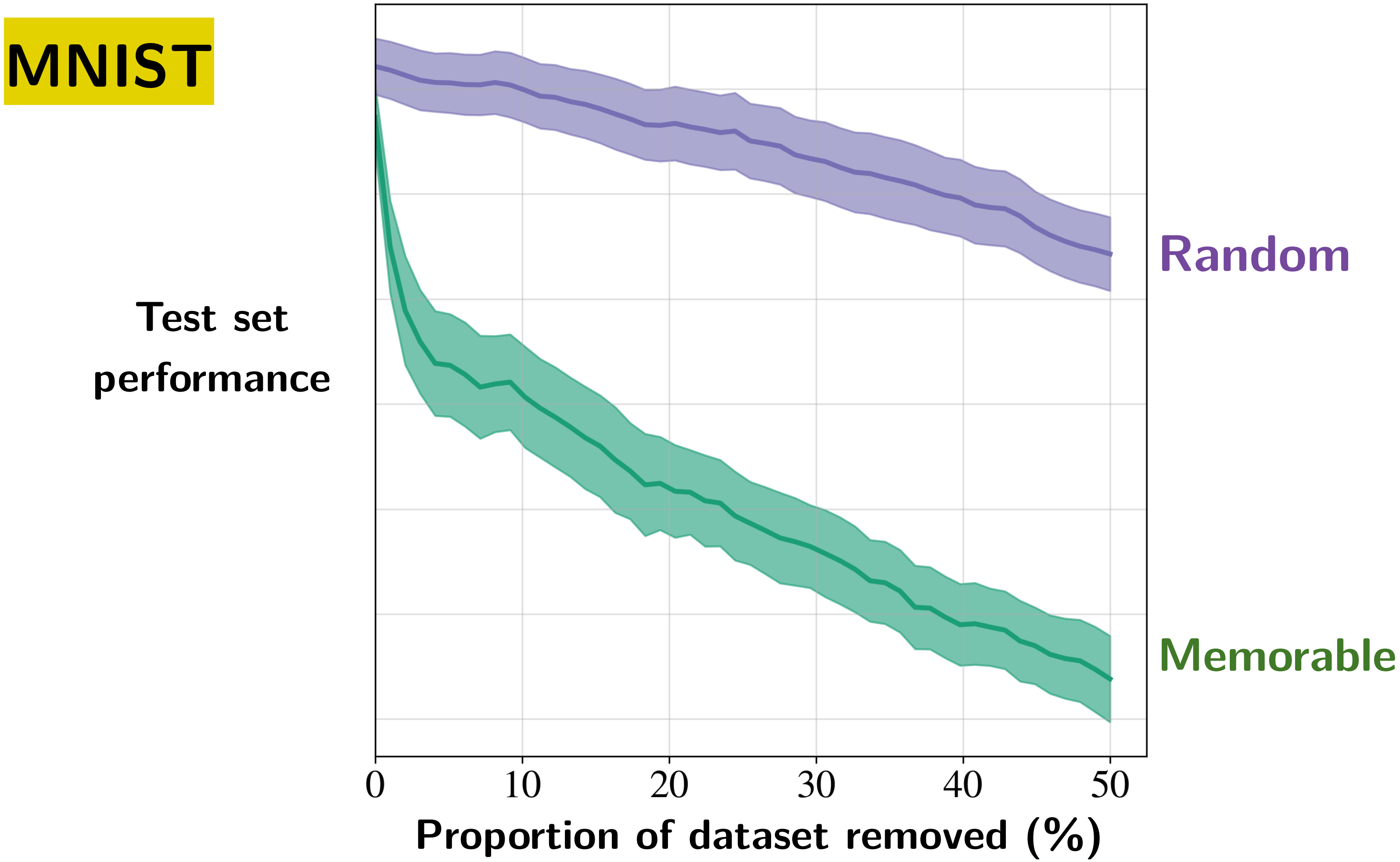
Most memorable **Least memorable**



Pan, P., et. al. **Continual deep learning by functional regularisation of memorable past.** NeurIPS, 2020.

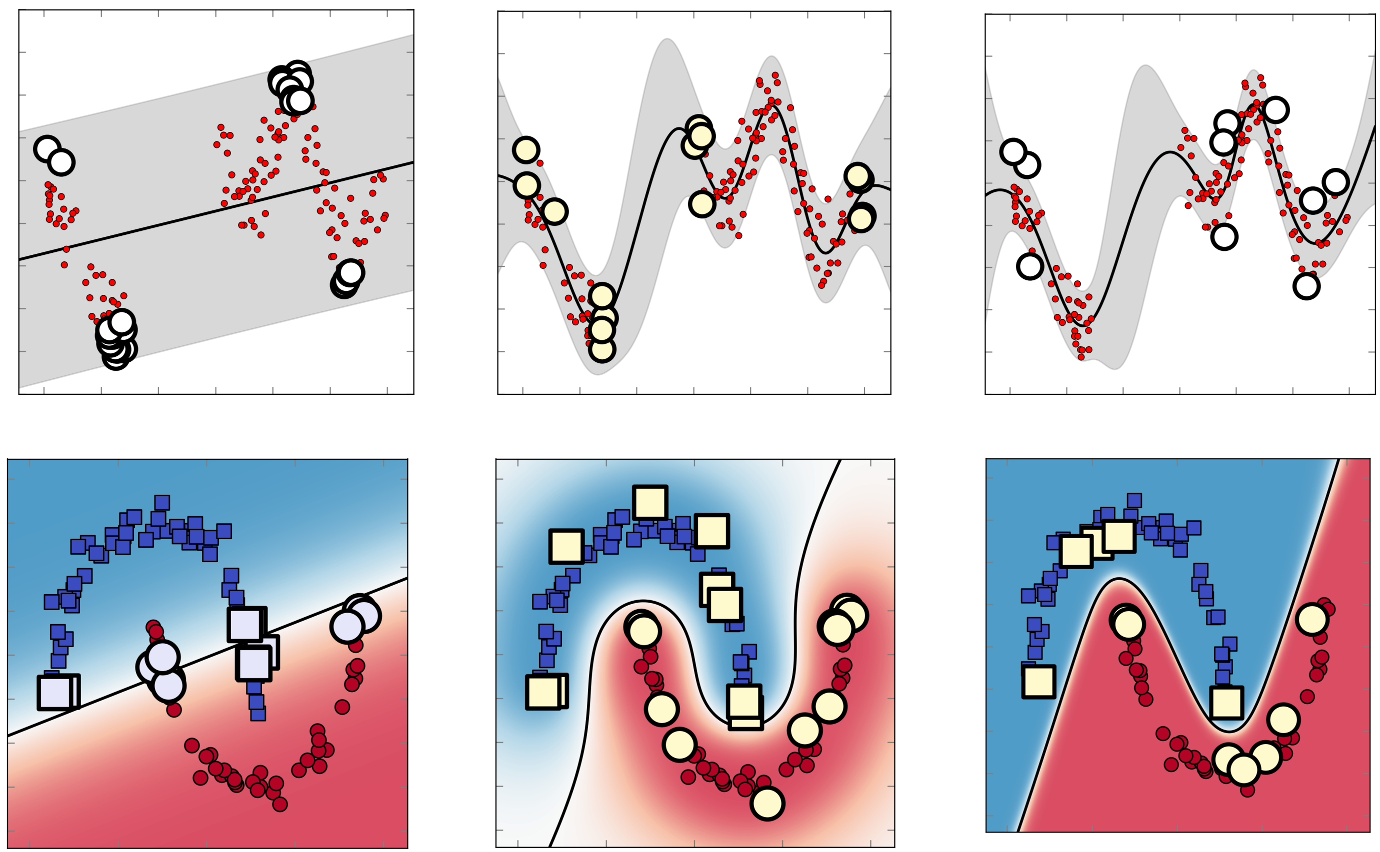
Memory Damage

- Memorable examples are the most impactful to model performance
- Demonstrated by removing examples in order of most to least memorable, retraining from scratch and evaluating the model on a fixed test set.



Conclusion

- The memorable experiences of a variety of machine-learning models can be identified with a **single Bayesian principle**.



Paper coming soon!