IBM **Research**

Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution

Tsuyoshi (Ide-san) Ide, Naoki Abe
{tide, nabe}@us.ibm.com
IBM T. J. Watson Research Center

Tsuyoshi Idé, Naoki Abe, "Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution," In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023, August 6-10, 2023, Long Beach, California, USA), pp. 845-856.

Agenda

- What is the task, "Anomaly Attribution"?
- What's wrong with the existing attribution methods?
- What is the new idea?
- Illustrative examples

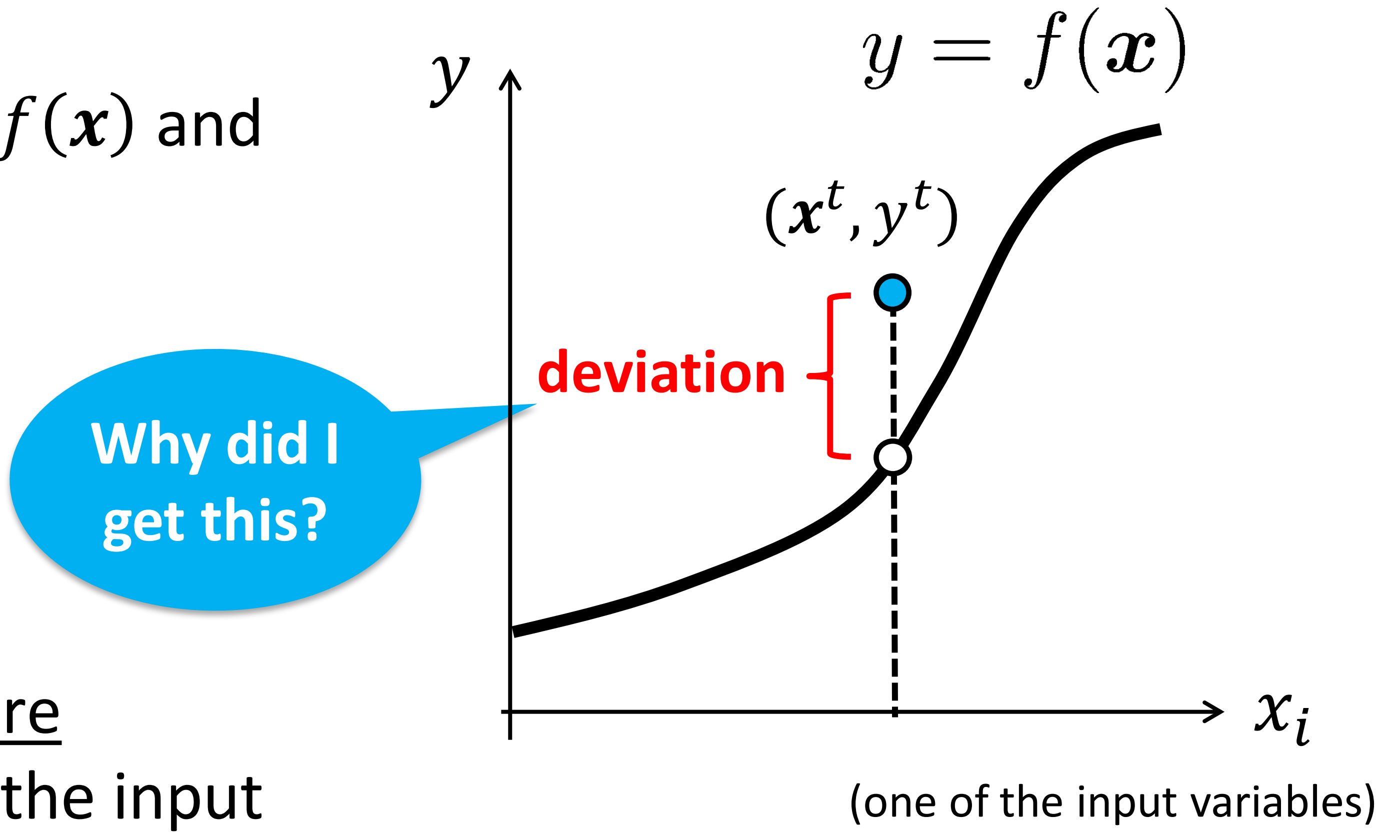
“Anomaly attribution” is an important topic in XAI (explainable AI) research.

Given:

- Black-box regression model $y = f(\boldsymbol{x})$ and a (set of) test sample $(\boldsymbol{x}^t, y^t)$
  - No access to the model beyond API
  - No access to the training data

Explain:

- The deviation $f(\boldsymbol{x}^t) - y^t$
- by computing the <u>attribution score</u> (responsibility score) for *each* of the input variables $\boldsymbol{x}$.



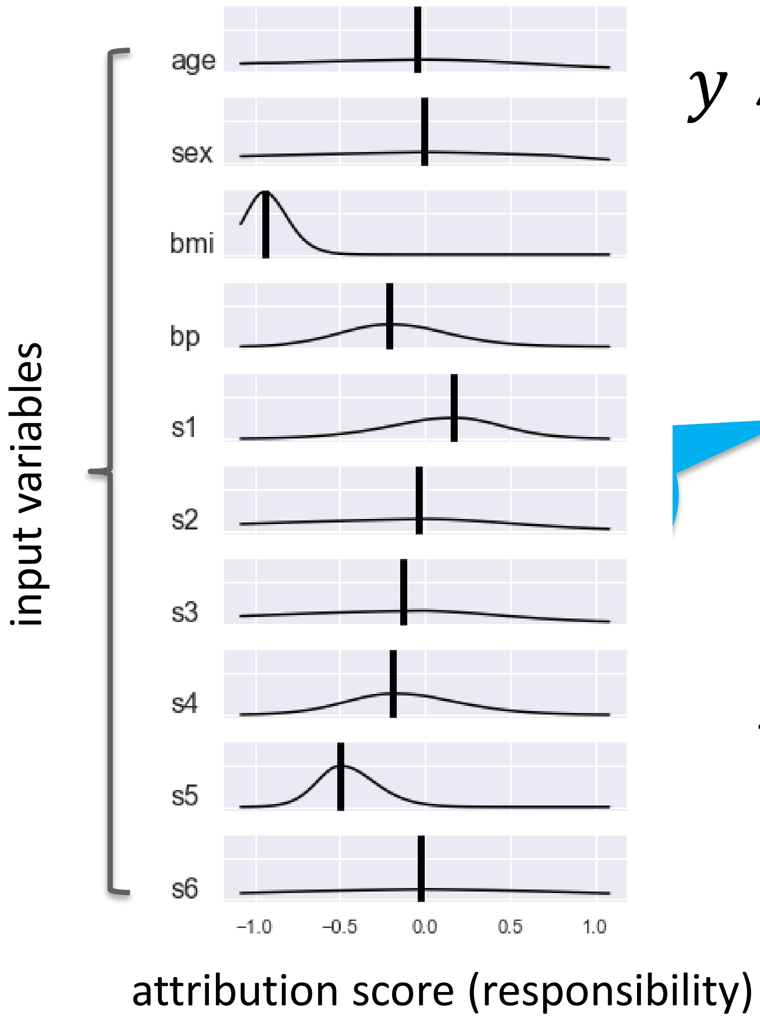
“Anomaly attribution” is an important topic in XAI (explainable AI) research.

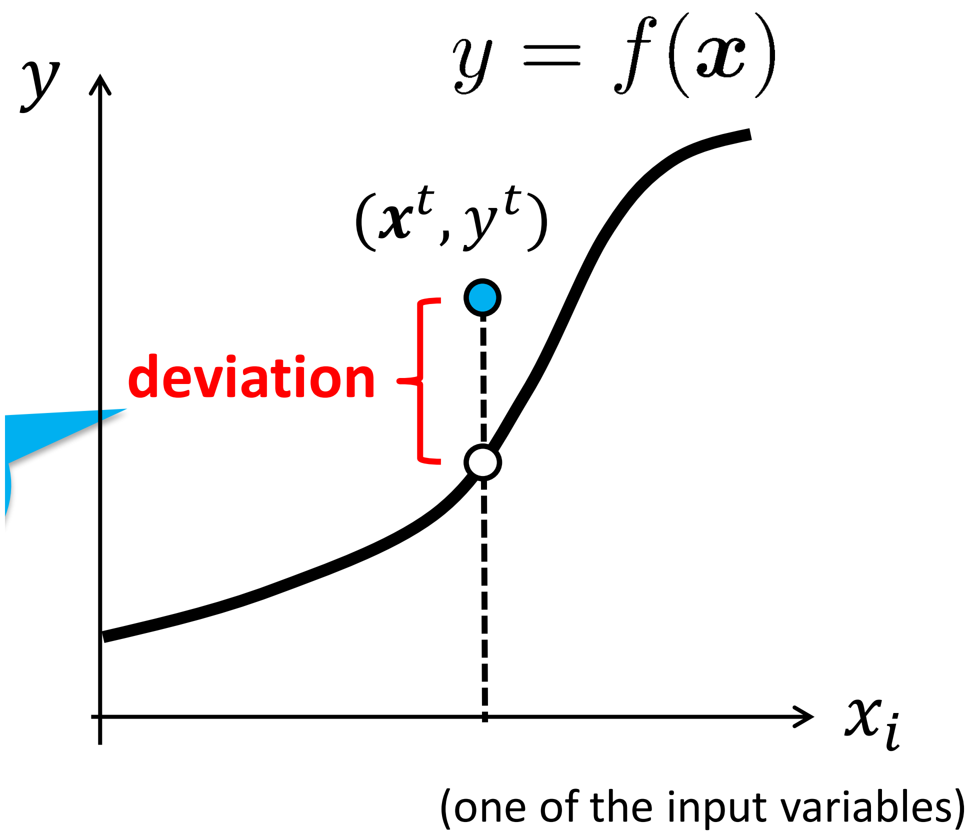
New capabilities that GPA has enabled

GPA = generative perturbation analysis (proposed method)

deviation-sensitive

probabilistic





Agenda

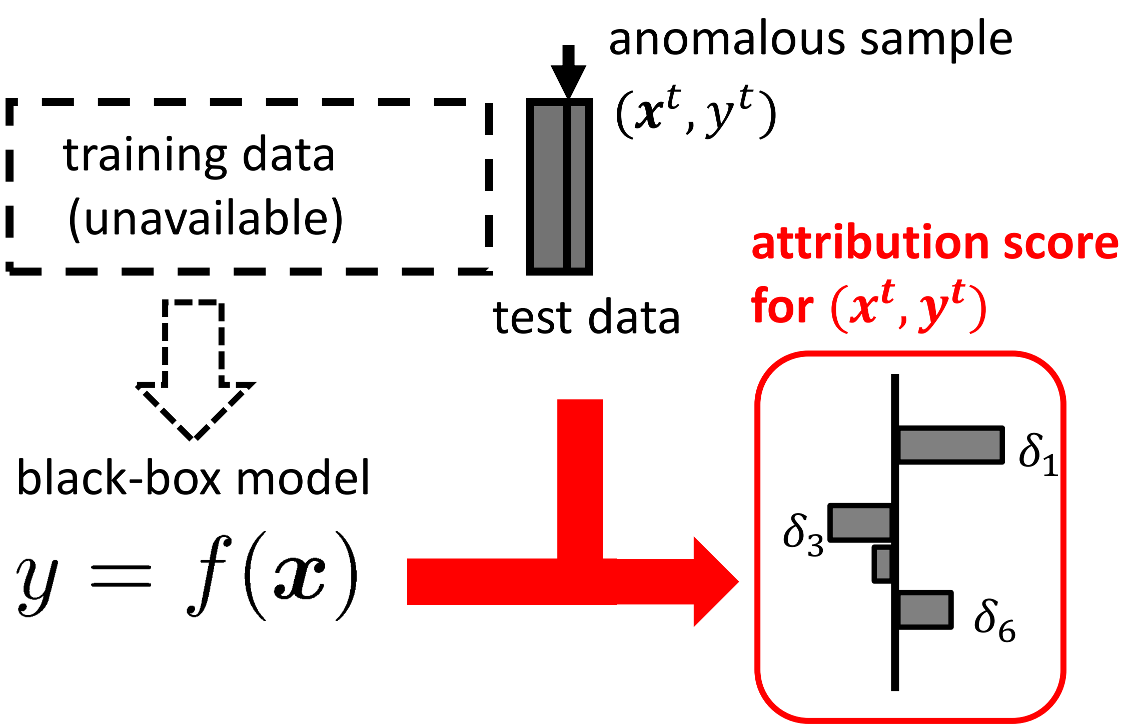
- What is the task, “Anomaly Attribution”?
- What’s wrong with the existing attribution methods?
- What is the new idea?
- Illustrative examples

LIME, Shapley values (SV), and integrated gradient (IG) are three major existing black-box attribution methods.

- LIME, SV, IG have the same in/output
  - In: black-box $y = f(\boldsymbol{x})$ and test sample.
  - Out: attribution score for each variable
- Why bother to develop a new method?

They are deviation-agnostic.

They can't compute score's uncertainty



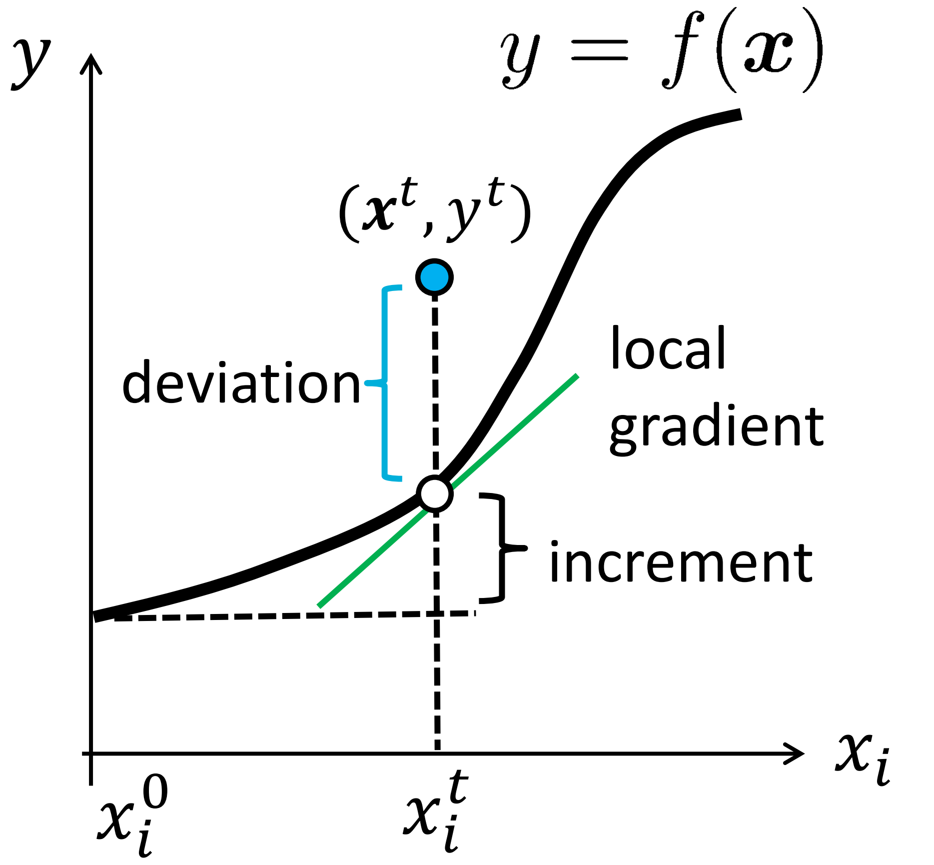
LIME, SV, and IG are to explain a black-box function itself locally.

- LIME = local gradient at $\boldsymbol{x}^t$
  - Gradient is numerically estimated via sampling.
- IG = increment from a reference point $\boldsymbol{x}^0$

$$\mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0) \triangleq (x_i^t - x_i^0) \int_0^1 \mathrm{d}\alpha \ \left. \frac{\partial f}{\partial x_i} \right|_{\boldsymbol{x}^0 + (\boldsymbol{x}^t - \boldsymbol{x}^0)\alpha}$$

- EIG = expected IG
  - Computed by marginalizing $\boldsymbol{x}^0$
- SV = (something mysterious)

$$\mathrm{SV}_i(\boldsymbol{x}^t) = \frac{1}{M} \sum_{k=0}^{M-1} \binom{M-1}{k}^{-1} \sum_{\mathcal{S}_i : |\mathcal{S}_i| = k} \left[\langle f \mid x_i^t, \boldsymbol{x}_{\mathcal{S}_i}^t \rangle - \langle f \mid \boldsymbol{x}_{\mathcal{S}_i}^t \rangle \right]$$



(one of the input variables)

Can they be used to explain deviations? – No.
Summary of theoretical results.

- Result 1: LIME, SV, IG, and EIG are deviation-agnostic
  - This is obvious from the original definition.
    - ✓ They explain $f(\boldsymbol{x})$ locally at $\boldsymbol{x} = \boldsymbol{x}^t$, independently y.
  - The conclusion still holds even when the target function is $f(\boldsymbol{x}) - y$ rather than $f(\boldsymbol{x})$.

- Result 2: SV is equivalent to EIG up to the second order of power expansion.

$$\mathrm{SV}_i(\boldsymbol{x}^t, y^t) \approx \mathrm{EIG}_i(\boldsymbol{x}^t, y^t)$$

- Result 3: LIME is equivalent to the derivative of IG and EIG

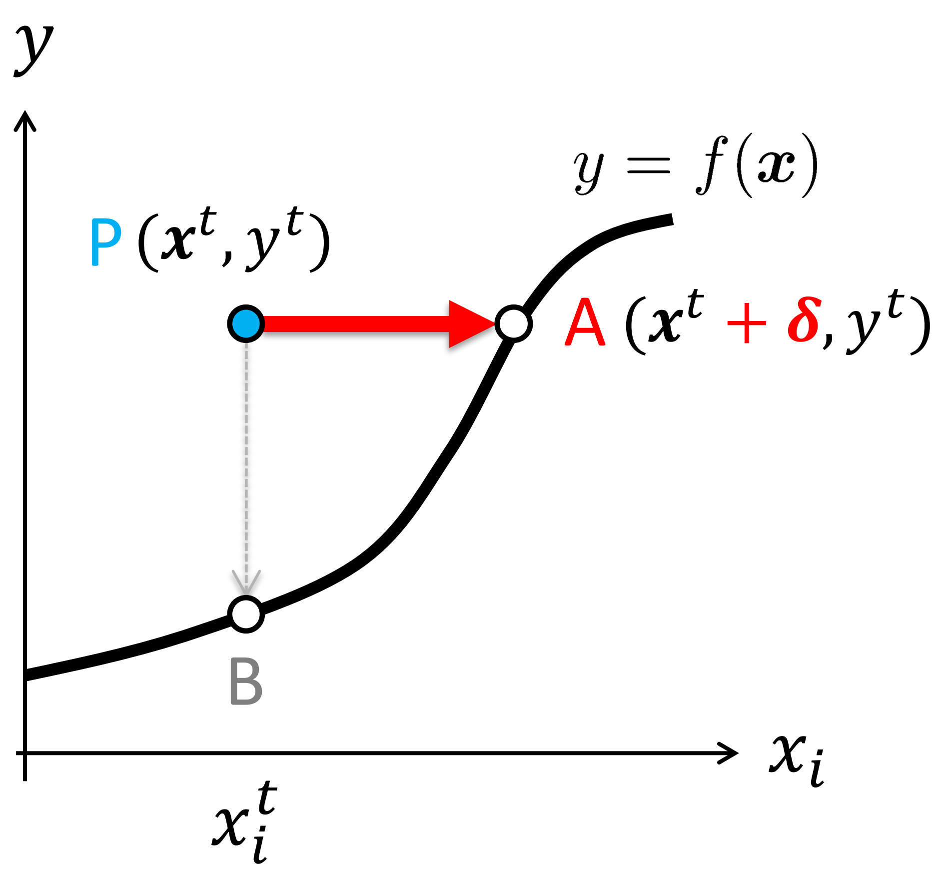
$$\mathrm{LIME}_i(\boldsymbol{x}^t, y^t) = \frac{\partial \mathrm{EIG}_i(\boldsymbol{x}^t, y^t)}{\partial x_i}$$

Agenda

- What is the task, “Anomaly Attribution”?
- What’s wrong with the existing attribution methods?
- What is the new idea?
- Illustrative examples

Given a test point (x^t, y^t) being anomalous, we ask: How much "work" would we need to bring it to the normalcy?

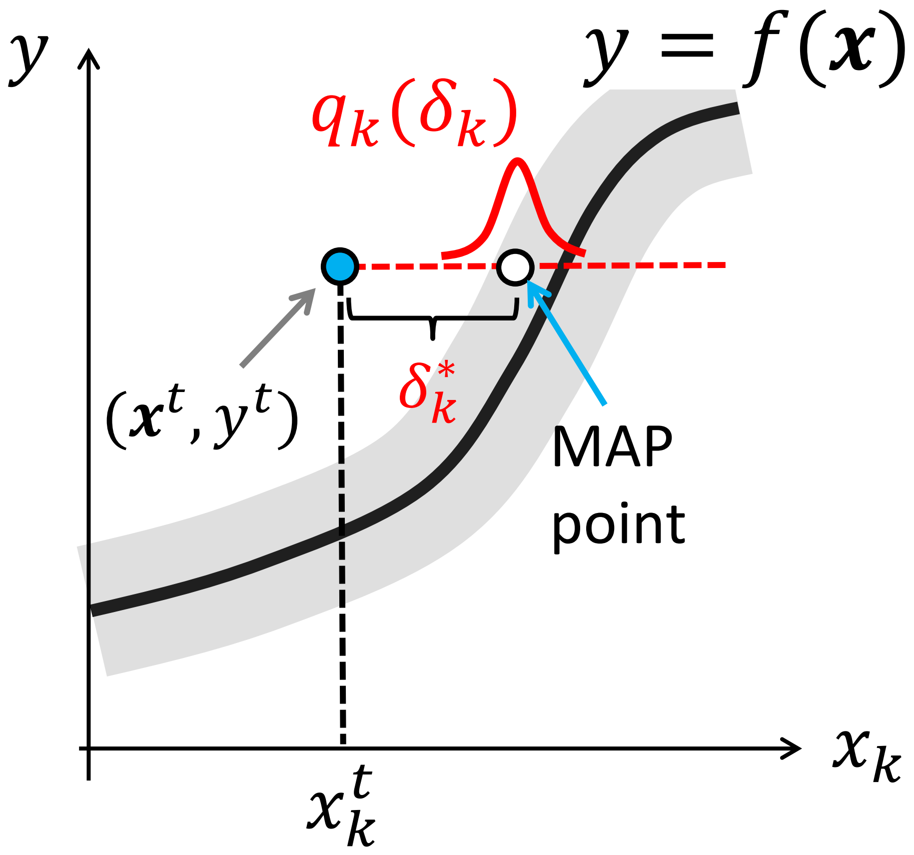
- The "work" required for each variable should be a natural attribution score.
- The outlier P wouldn't have been anomalous if it were at A.
- Hence, the amount of shift, $\boldsymbol{\delta}$, can be viewed as the "work," indicating the responsibility of each variable.
- How about B? We need a help of $p(y \mid \boldsymbol{x})$.



Perturbation as explanation: Our goal is to find the posterior distribution of δ.

- We need a generative model to handle the ambiguity in prediction.
  - The on-the-curve points may not represent normalcy.
- Generative process with $\boldsymbol{\delta}$ as model parameter.
  - $p(y \mid \boldsymbol{x}, \boldsymbol{\delta}, \lambda) = \mathcal{N}(y \mid f(\boldsymbol{x} + \boldsymbol{\delta}), \lambda^{-1})$
  - priors (η, a_0, b_0 are hyperparameters):
    - ✓ $p(\boldsymbol{\delta}) = \mathcal{N}(\boldsymbol{\delta} \mid \mathbf{0}, \eta \mathbf{I})$
    - ✓ $p(\lambda) = \mathrm{Gam}(\lambda \mid a_0, b_0)$
- Formal solution of the posterior (typically $N_{\text{test}} = 1$)

$$Q(\boldsymbol{\delta}) \propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\text{test}}} \int_0^\infty \mathrm{d}\lambda \; p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) p(\lambda)$$

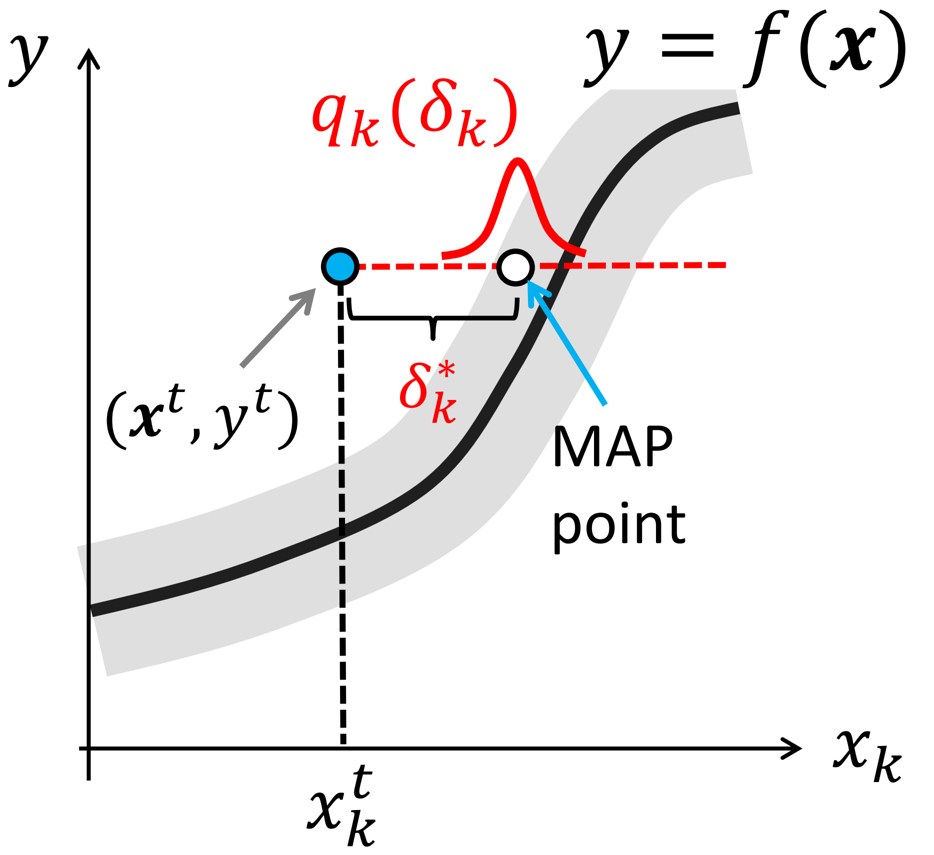


Using variational Bayesian approach combined with a mean-field-like approximation to get variable-wise posteriors.

- Formal solution of the posterior (typically $N_{\mathrm{test}} = 1$)

$$Q(\boldsymbol{\delta}) \propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\mathrm{test}}} \int_0^\infty \mathrm{d}\lambda \; p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) p(\lambda),$$

$$\propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\mathrm{test}}} \frac{1}{\sqrt{b_0}} \left\{ 1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b_0} \right\}^{-(a_0 + \frac{1}{2})},$$



- How do we get a variable-wise distribution?
  - We find an approximated solution by minimizing the KL divergence between $Q(\boldsymbol{\delta})$ and a factorized from:

$$Q(\boldsymbol{\delta}) = Q(\delta_1, \ldots, \delta_M) \approx \prod_{k=1}^{M} q_k(\delta_k),$$

  - We also use a mean-field-like approximation to get an explicit form of $\{q_k(\delta_k)\}$. → paper

(For ref.) How the GPA algorithm works

- GPA algorithm has two parts:
  - MAP (maximum a posteriori) estimation
  - Distribution estimation
- MAP estimation solves:

$$\min_{\boldsymbol{\delta}} \left\{ \frac{\eta}{2}\|\boldsymbol{\delta}\|_2^2 + \ln\left\{1 + \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2b(\boldsymbol{x}^t)}\right\}^{\frac{2a_0+1}{2}} \right\}$$

  - Use proximal gradient (with ℓ_1 regularizer)
  - The gradient is estimated via local sampling (like LIME)
- Distribution estimation uses a mean-field approximation
  - "Think of the others fixed to the MAP value and focus on yourself."

Algorithm 2 Generative Perturbation Analysis

Require: $f(\boldsymbol{x})$, $\mathcal{D}_{\text{test}}$, parameters $\eta, \nu, \kappa, a_0, \{b(\boldsymbol{x}^t)\}$.

randomly initialize $\boldsymbol{\delta} \approx \mathbf{0}$.

MAP

repeat
  set $\boldsymbol{g} = 0$
  for all $(y^t, \boldsymbol{x}^t) \in \mathcal{D}_{\text{test}}$ **do**
    Compute the local gradient $\frac{\partial f(\boldsymbol{x}^t+\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}$
    Update $\boldsymbol{g} \leftarrow \boldsymbol{g} + \frac{\partial f(\boldsymbol{x}^t+\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \frac{y^t - f(\boldsymbol{x}^t+\boldsymbol{\delta})}{2b(\boldsymbol{x}^t)+[y^t-f(\boldsymbol{x}^t+\boldsymbol{\delta})]^2}$
  end for
  $\boldsymbol{g} \leftarrow (1-\kappa\eta)\boldsymbol{\delta} + \kappa(2a_0+1)\boldsymbol{g}$
  $\boldsymbol{\delta} = \text{sign}(\boldsymbol{g}) \max\{0, |\boldsymbol{g}| - \eta\nu\}$
until convergence

set $\boldsymbol{\delta}^* = \boldsymbol{\delta}$

distribution

for all k **do**
  $q_k(\delta) = Q(\delta_1^*, \ldots, \delta_{k-1}^*, \delta, \delta_{k+1}^*, \delta, \delta_M^*)$
  $q_k(\cdot) \leftarrow q_k(\cdot)/\int d\delta' q_k(\delta')$ with Eq. (18)
end for

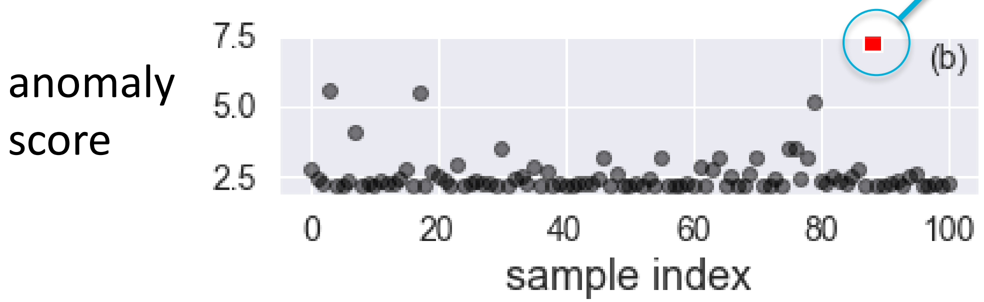
return $\{q_k(\cdot) \mid k = 1, \ldots, M\}$ and $\boldsymbol{\delta}^*$

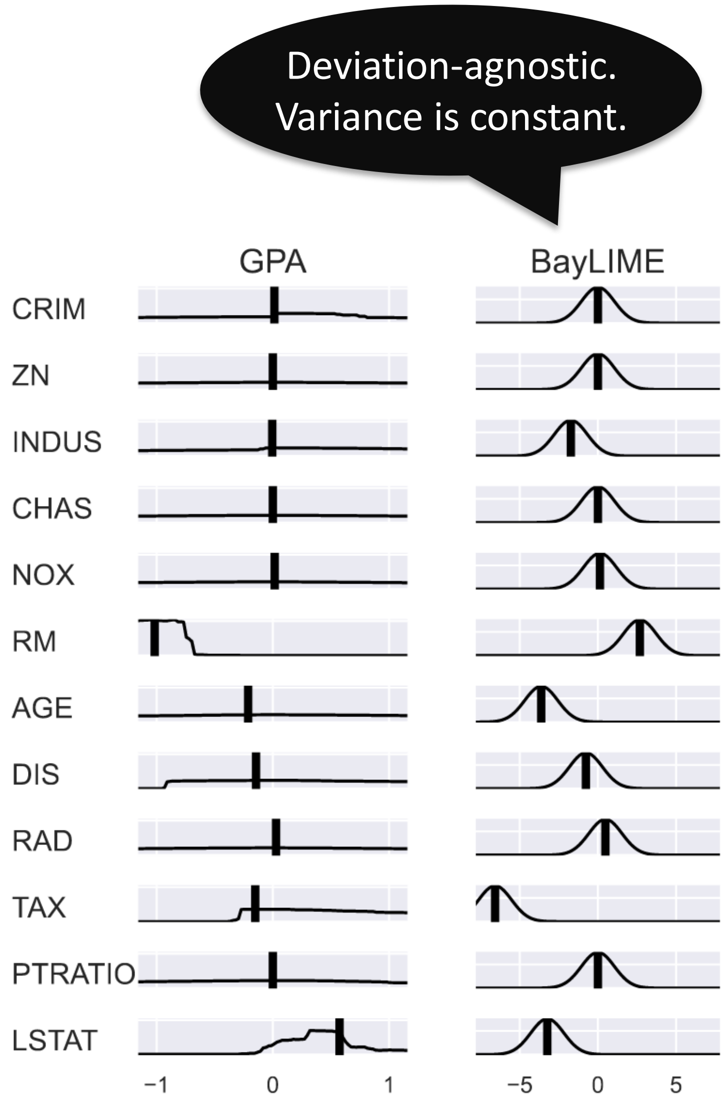
Agenda

- What is the task, “Anomaly Attribution”?
- What’s wrong with the existing attribution methods?
- What is the new idea?
- Illustrative examples

"Why does this house look so unusual?"
House hunting use-case

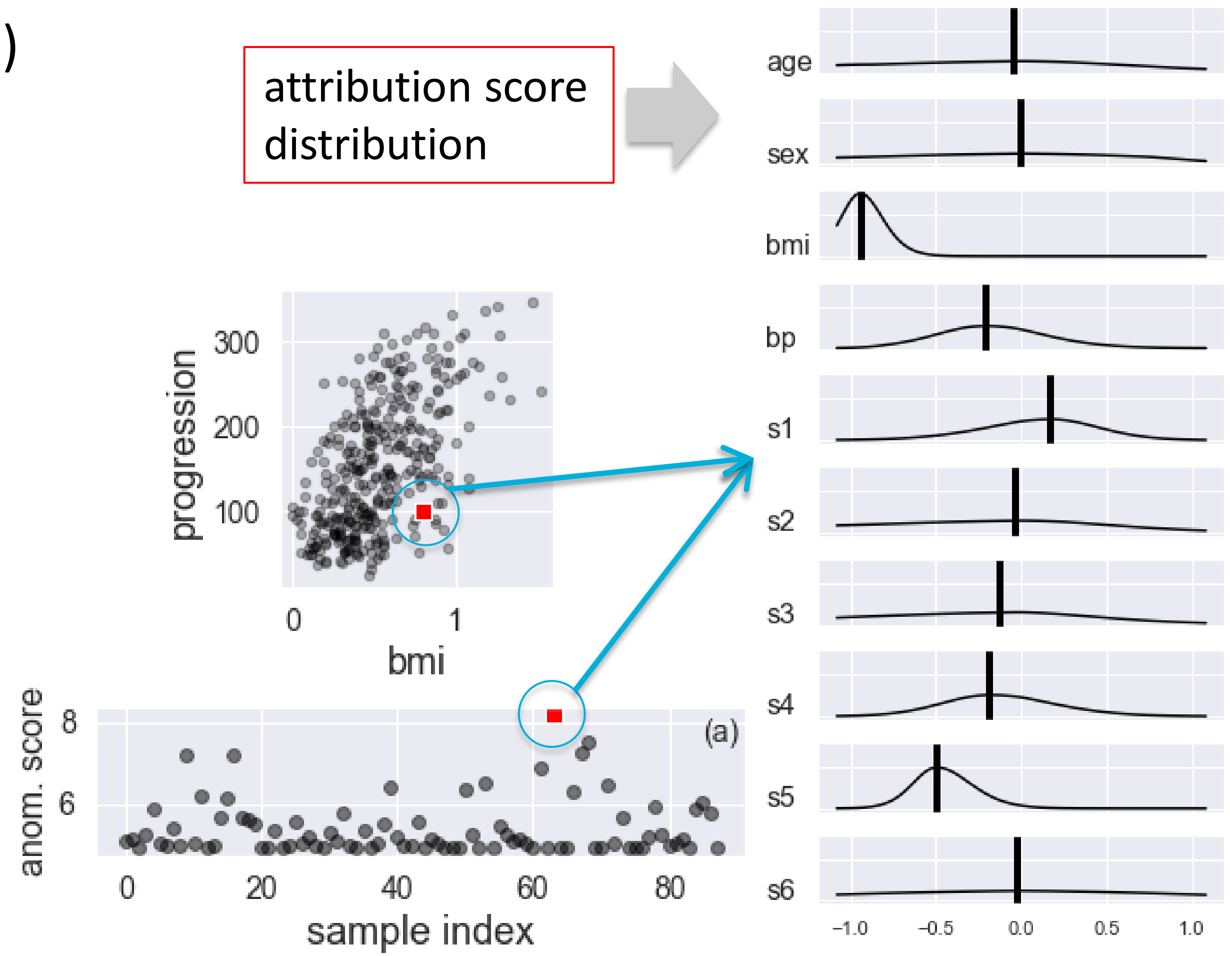
- Boston Housing data
  - y: house price
  - **x**: house age, # rooms, neighborhood crime rate, etc.
- Computed attribution scores for the top outlier.
  - GPA is able to provide variable-specific distributions in contrast to BayLIME
- Is it a bargain? Probably yes.
  - It's got unusually larger #rooms (RM) and lower poor neighbors (LSTAT) than the peers in the same price range.





"Why does this patient look so unusual?"
Healthcare use-case

- Diabetes data
  - y: diabetes' progression (numerical score)
  - **x**: biomarkers (BMI, blood pressure, etc.)
- Computed attribution score for the top outlier (patient # 63).
  - Found a large negative score in BMI
    - ✓ The high and narrow pdf translates to high confidence
  - For his progression level, he would look like a regular patient if BMI were much smaller:
    - ✓ "He is overweight but healthy (low progression)" or "He is healthy despite overweight"



Summary

- GPA is the first black-box attribution framework allowing probabilistic attribution.
- We have showed a strong impossibility result: LIME, SV, and IG are deviation-agnostic, and hence, not suitable for anomaly attribution.
- We have also uncovered a relationship between LIME, SV, and IG for the first time.

	model-agnostic	training-data-free	baseline-input-free	y-sensitive	built-in UQ
LIME [34]	yes	yes	yes	no	yes/no
SV [43,44]	yes	no	yes	no	no
IG [39,46]	yes	yes	no	no	no
EIG [6]	yes	no	yes	no	no
Z-score [5]	yes	no	yes	no	no
LC [20]	yes	yes	yes	yes	no
GPA	**yes**	**yes**	**yes**	**yes**	**yes**

Thank you!