IBM **Research**

Attributing anomalies from black-box predictions

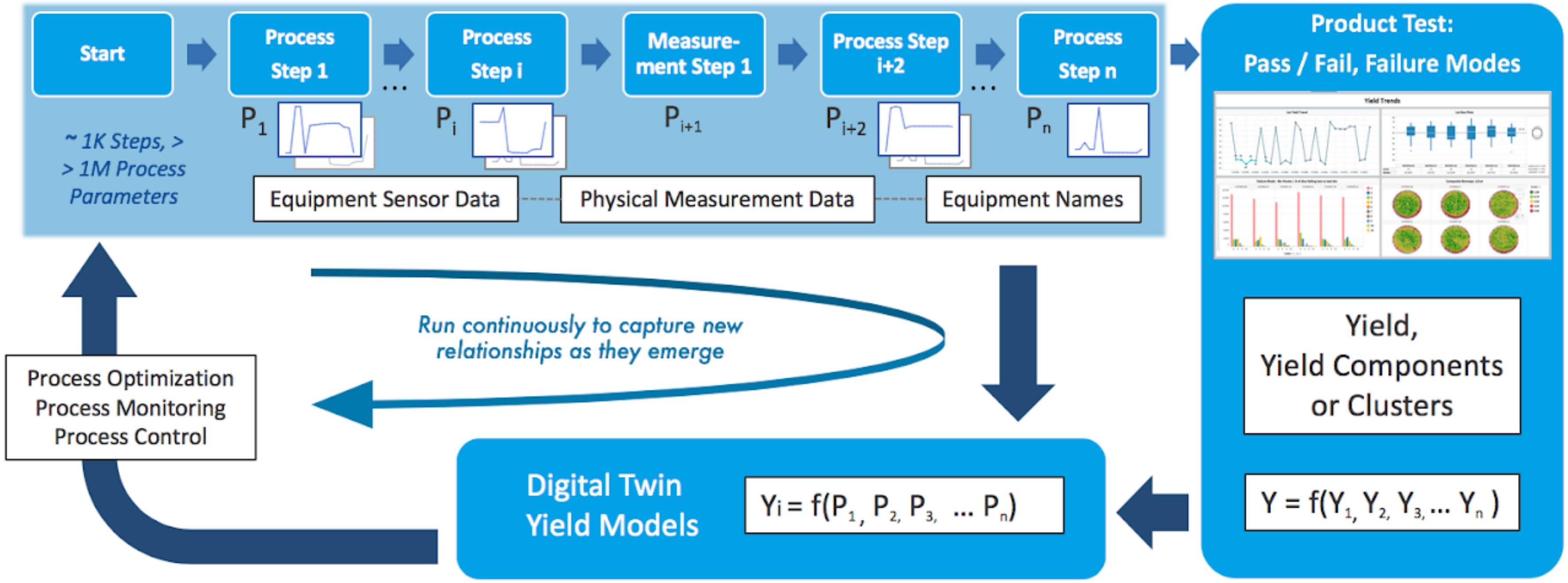
Tsuyoshi (Ide-san) Ide (井手 剛)
tide@us.ibm.com
IBM Semiconductors, T. J. Watson Research Center

Agenda

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

Digital twin is a black-box function to predict a KPI. Explainability is crucial.

- Example: Yield prediction as a function of process parameters.
  - Mfg. process is so complex that data-driven models (e.g., DNN) are used to get $y = f(\cdot)$.



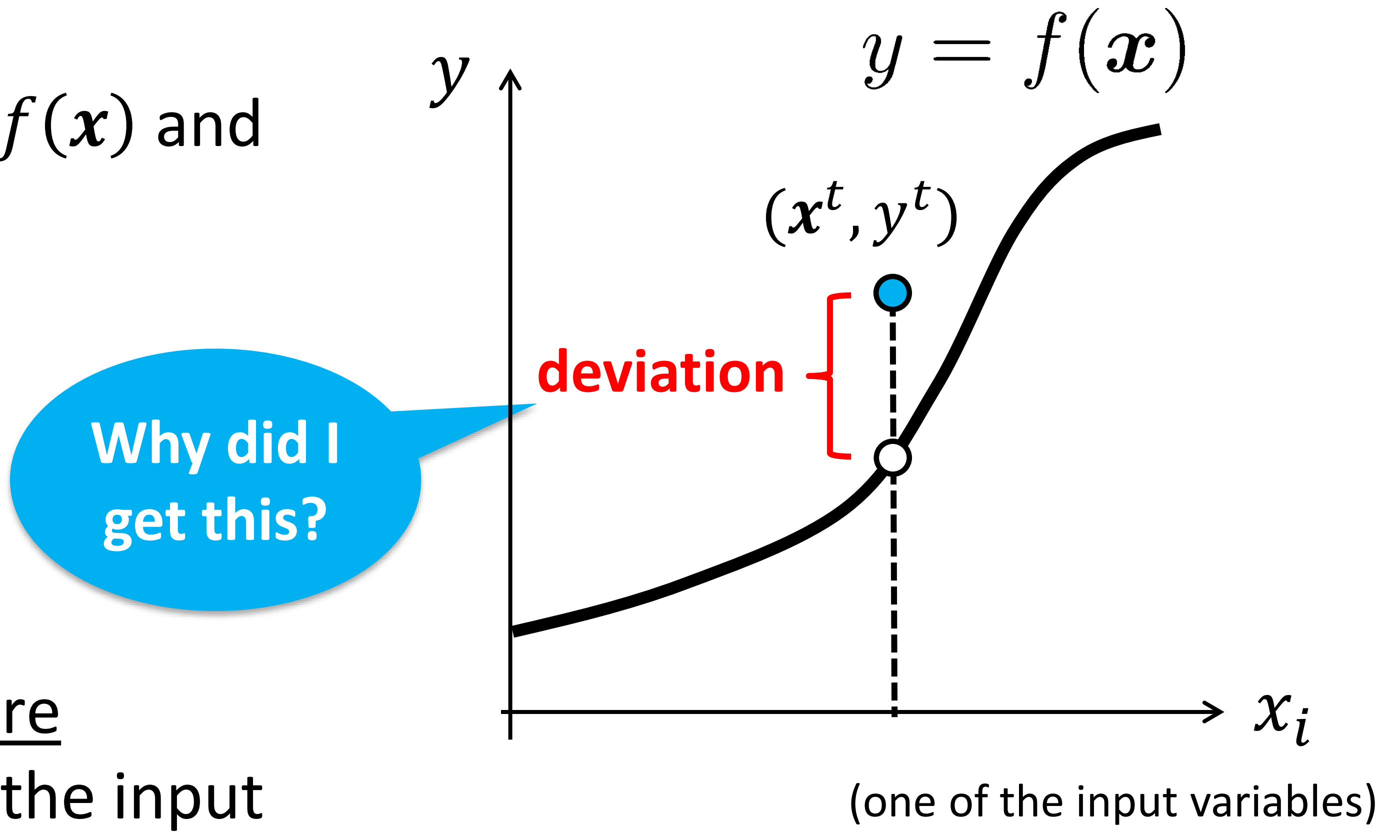
- Explainability of prediction is critical for process improvement

Sam Seto, TIBCO Community Article, https://community.tibco.com/s/article/digital-twins-yield-wide-data-manufacturing-using-data-function-tibcor-data-science-team

“Anomaly attribution” addresses the key question of digital twins

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



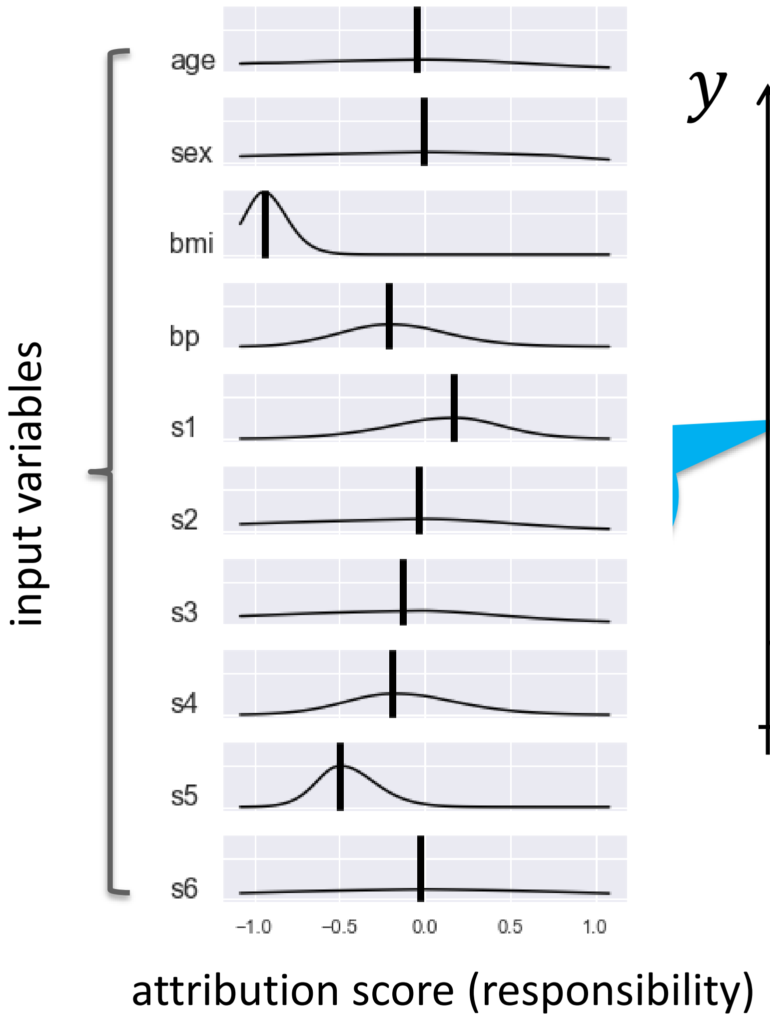
Explain:

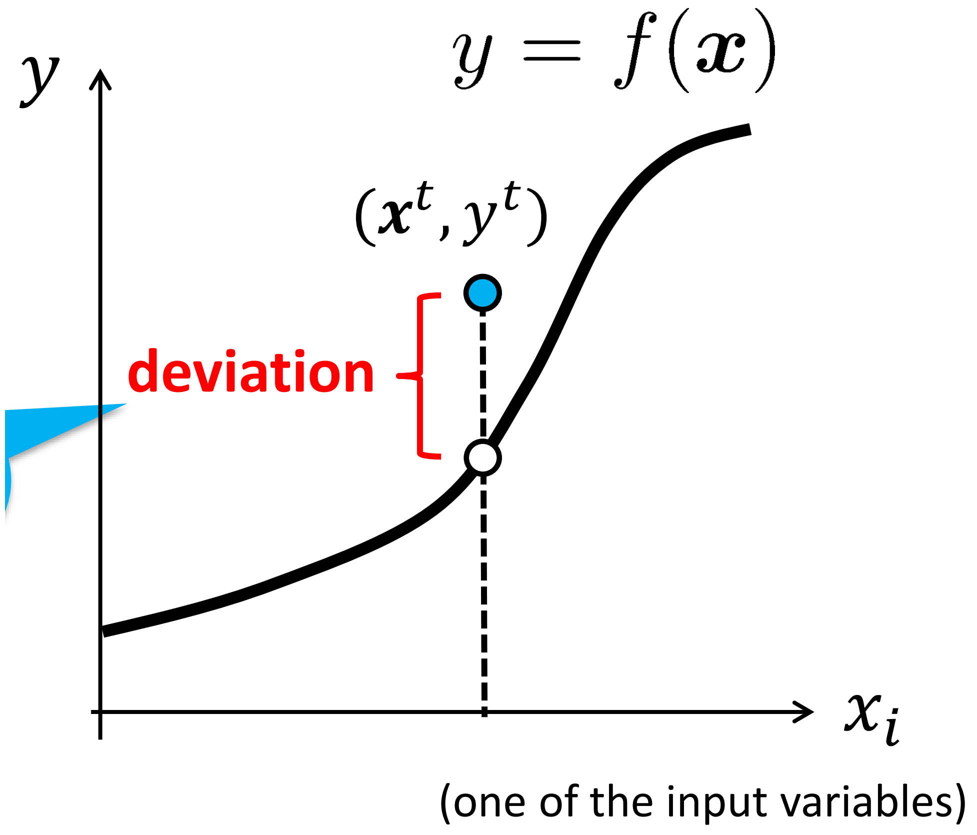
- The deviation $f(\boldsymbol{x}^t) - y^t$
- by computing the attribution score (responsibility score) for *each* of the input variables $\boldsymbol{x}$.

Seeking an automated way of computing the responsibility of an observed anomaly

Practical requirements of anomaly attribution

- Able to explain the deviation (Sounds obvious, huh?)
- Able to compute the uncertainty of the score (challenging)





Agenda

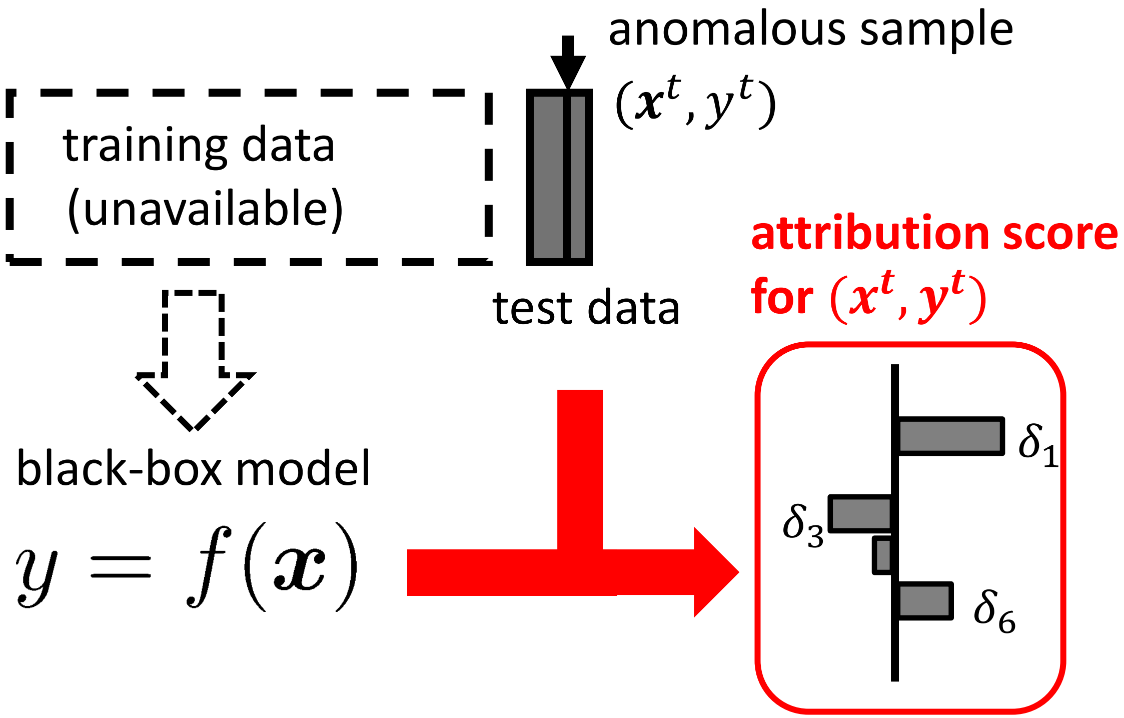
- What is the task, “Anomaly Attribution”?
- What’s wrong with the existing attribution methods?
- What is the new idea?
- Illustrative examples

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

- LIME, SV, IG are well-established model-agnostic attribution methods
  - In: black-box $y = f(\boldsymbol{x})$ and test sample.
  - Out: attribution score for each variable
- Why bother to develop a new method?

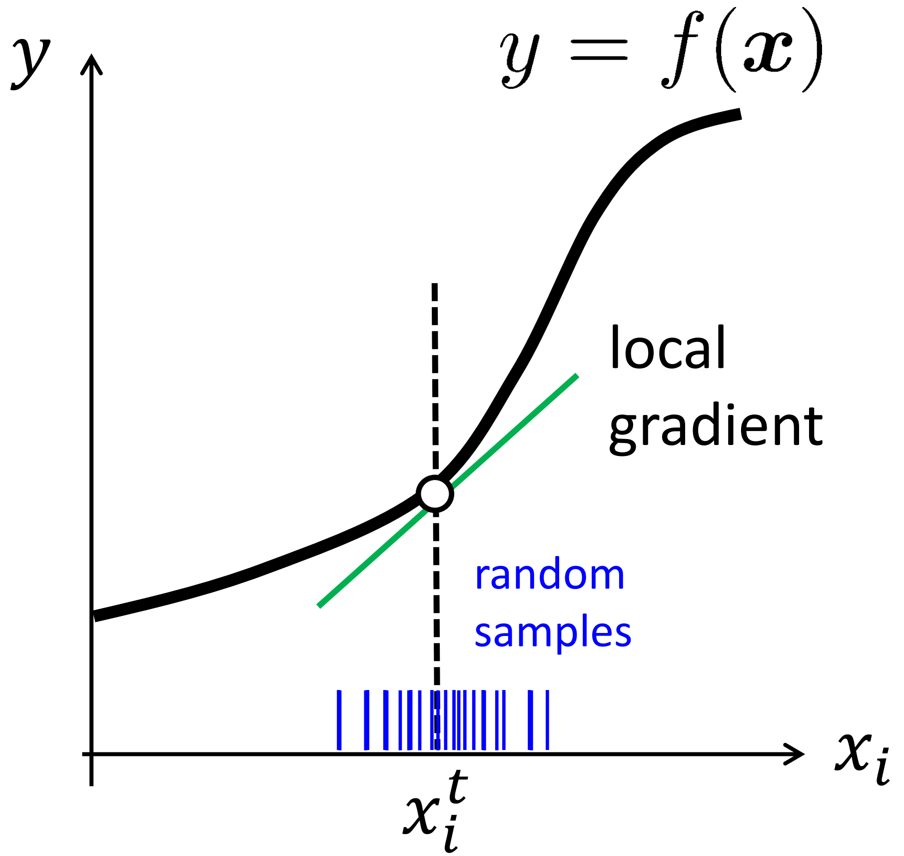
They are, in fact, deviation-agnostic.

They can't compute score's uncertainty



(For ref.) LIME [Ribeiro+ 16] does local sensitivity analysis of the black-box function

- Sensitivity = gradient = attribution score
- Challenge:
  - $f(\boldsymbol{x})$ is black-box; No way of getting the gradient analytically.
- Idea:
  - Randomly generate samples around $\boldsymbol{x}^t$
    - ✓ $\{(\boldsymbol{x}^{t[1]}, y^{t[1]}), \dots, (\boldsymbol{x}^{t[1]}, y^{t[N]})\}$ where $y^{t[n]} = f(\boldsymbol{x}^{t[n]})$.
  - Fit a (sparse) linear model (lasso)
    - ✓ $y = \boldsymbol{a}^\top \boldsymbol{x} + b$
  - The regression coefficients is an estimator of the gradient (= explanation).



- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).
- LIME: Local Interpretable Model-agnostic Explanations

(For ref.) Integrated gradient (IG) computes the increment from a reference point

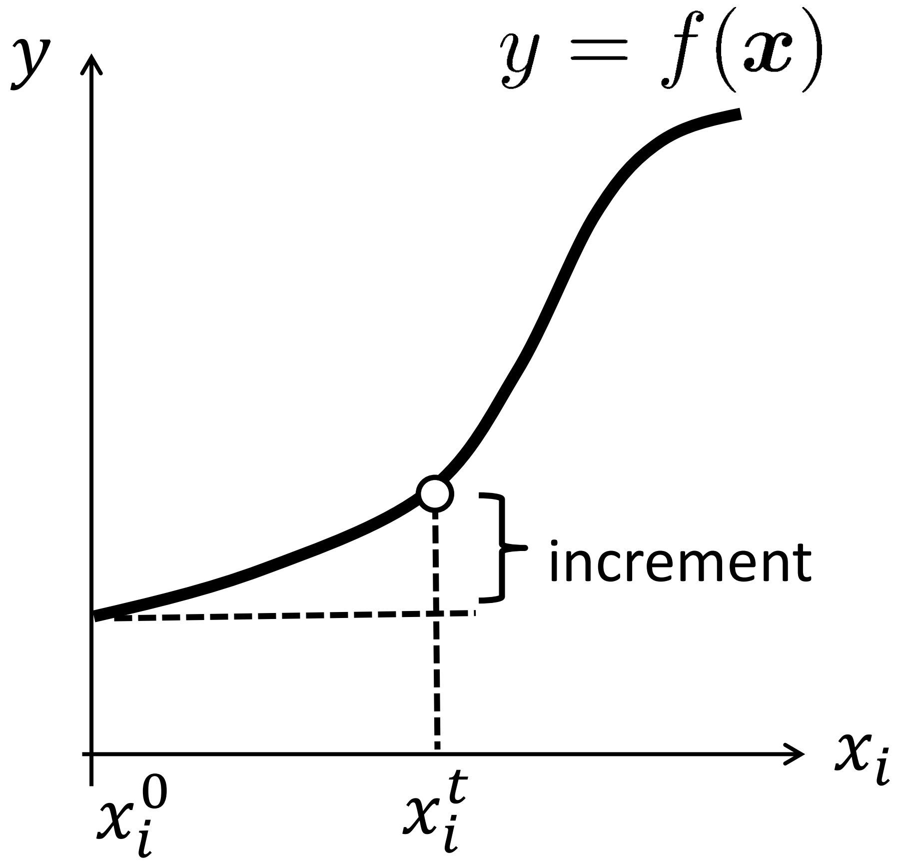
- Definition of IG [Sipple 20]
  - Increment from a reference point $\boldsymbol{x}^0$

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

  - The gradient is numerically estimated with a LIME-like approach.
  - The integral is also evaluated numerically
- Expected IG (EIG) [Deng+ 21]
  - Computed by marginalizing $\boldsymbol{x}^0$ with a distribution of the reference point

$$\mathrm{EIG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0) \triangleq \int \mathrm{d}\boldsymbol{x}^0 \, \underline{P(\boldsymbol{x}^0)} \mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0)$$

typically empirical distribution of the training samples



- John Sipple. "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure," In Proceedings of the 37th International Conference on Machine Learning (ICML 20).
- Huiqi Deng, et al. , A Unified Taylor Framework for Revisiting Attribution Methods. In Proceedings of the AAAI Conference on Artificial Intelligence. 11462–11469, 2021.

(For ref.) Shapley values (SV) originate from game theory.

- SV originated from game theory and are defined without relying on geometric interpretations.
- The definition is a bit nonintuitive:

$$\mathrm{SV}_i(\boldsymbol{x}^t) = \frac{1}{M}\sum_{k=0}^{M-1}\binom{M-1}{k}^{-1}\sum_{\mathcal{S}_i:|\mathcal{S}_i|=k}\Delta f(x_i^t, \mathcal{S}_i)]$$

  - S_i: A variable set (of size k) excluding i.
  - If $M = 4, i = 1, k = 2$, and $S_1 = \{2,3\}$,
    - ✓ $\Delta f(x_1^t, S_1) = \frac{1}{N}\sum_{n=1}^{N}[f\left(x_1^t, x_2^t, x_3^t, x_4^{(n)}\right) - f\left(x_1^{(n)}, x_2^t, x_3^t, x_4^{(n)}\right)]$

- SV quantifies the impact of the i-th variable by contrasting the expected values when x_i is set to x_i^t, versus when x_i is averaged out.

- SV looks mysterious, but fortunately (and unexpectedly), SV ≈ EIG holds!

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 41, 3 (2014), 647–665.

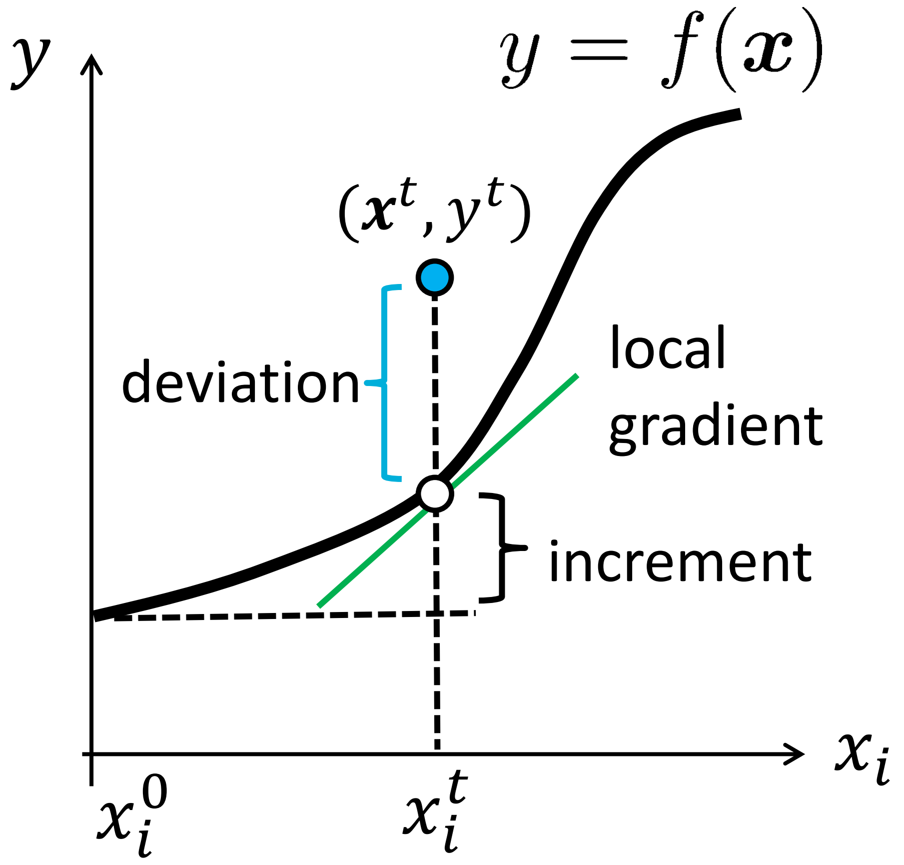
Can they explain deviations by changing the target to $f(x) - y$? – Actually, no. Summary of theoretical results [Ide-Abe 23].

- 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

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



(one of the input variables)

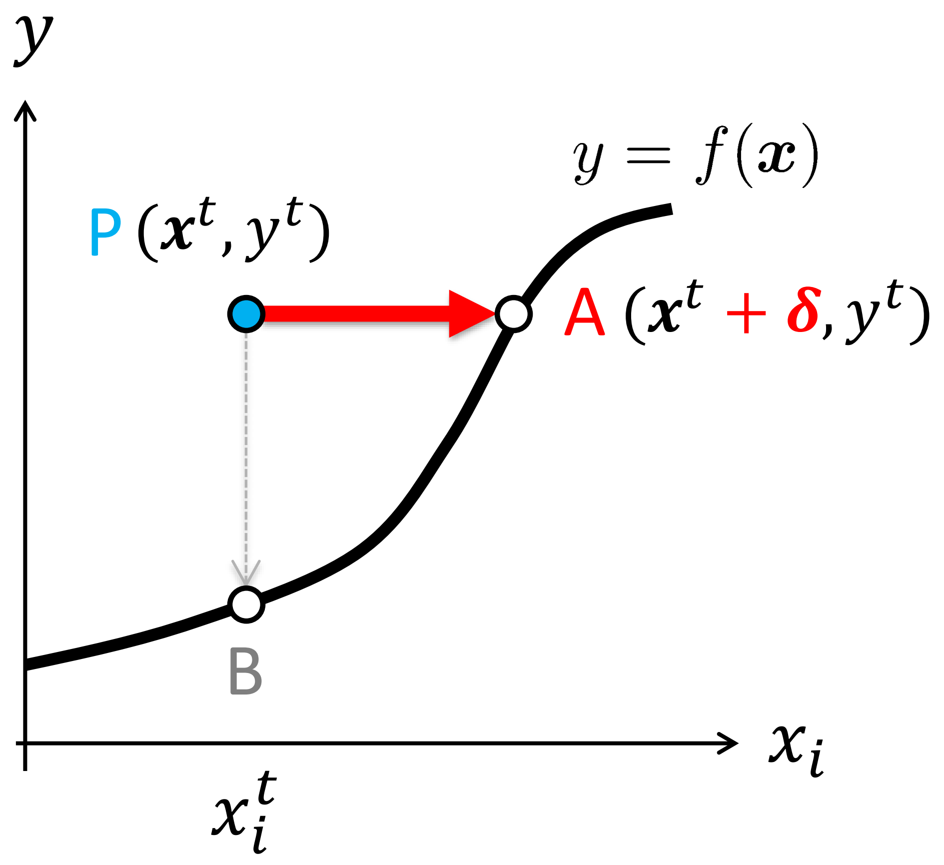
T. Idé, N Abe, "Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution," In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023, August 6-10, 2023, Long Beach, California, USA), pp. 845-856.

Agenda

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

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 $p(y \mid \boldsymbol{x})$.

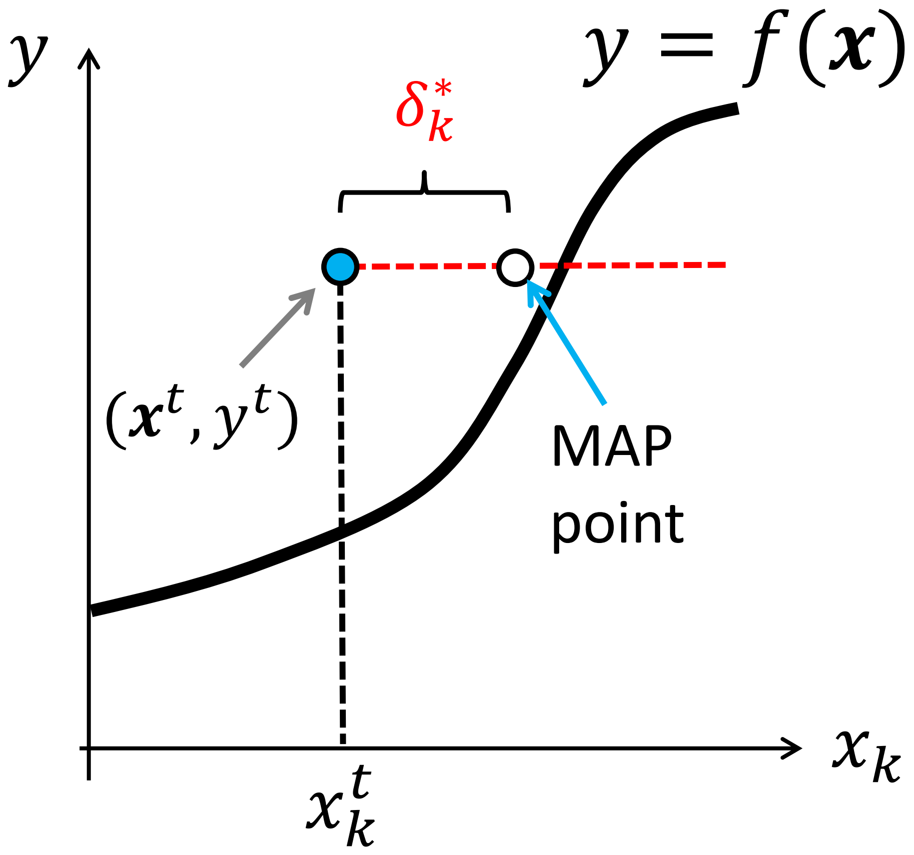


Perturbation as explanation: Likelihood compensation (LC) [Ide+ 21]

- 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.
  - observation: $p(\, y \mid \boldsymbol{x}, \boldsymbol{\delta}, \lambda \,) = \mathcal{N}(y \mid f(\boldsymbol{x} + \boldsymbol{\delta}), \lambda^{-1})$
  - prior: $p(\boldsymbol{\delta}) = \mathcal{N}(\, \boldsymbol{\delta} \mid \mathbf{0}, \eta \mathbf{I} \,)$

- $\boldsymbol{\delta}$ can be determined by solving
  - $\delta^* = \mathrm{argmax}_\delta \frac{1}{N_\mathrm{test}} \sum_{t=1}^{N_\mathrm{test}} \ln p(\, y^t \mid x^t, \boldsymbol{\delta}, \lambda \,) p(\boldsymbol{\delta})$
    - ✓ Typically, $N_\mathrm{test} = 1$



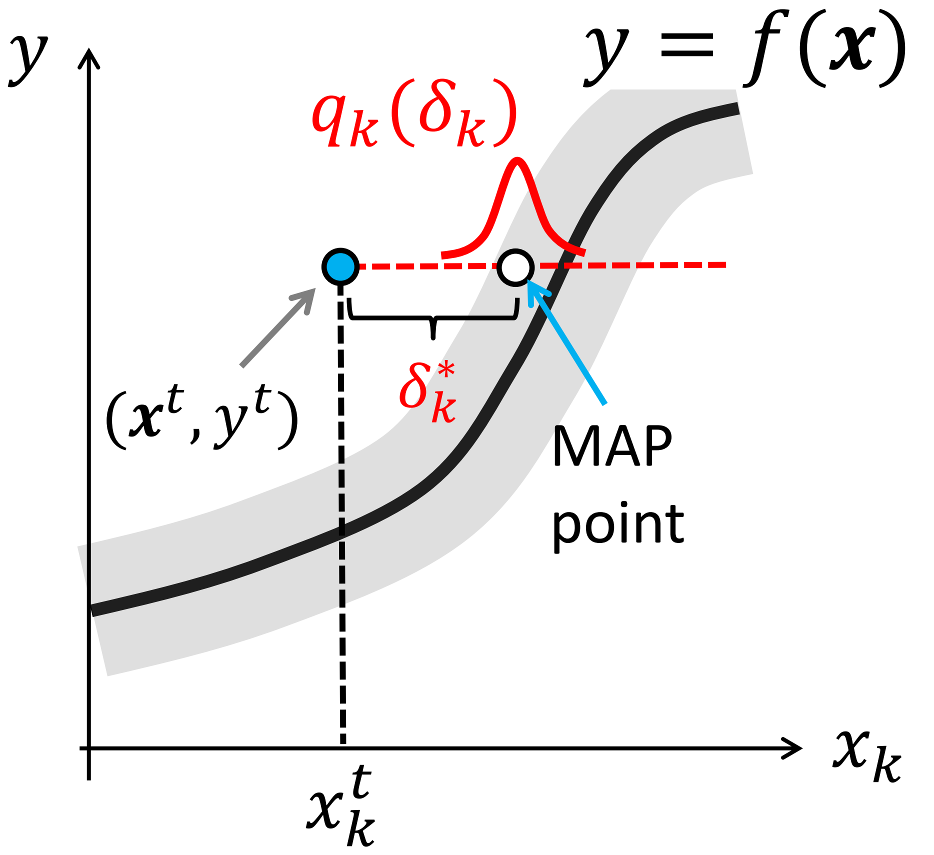
T. Idé, et al., Naoki Abe, “Anomaly Attribution with Likelihood Compensation,” In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 21, February 2-9, 2021, virtual), pp.4131-4138

Generative perturbation analysis (GPA) [Ide-Abe 23]: Extending LC to incorporate uncertainty quantification

- The generative process can be viewed as a Bayesian inference model for $\boldsymbol{\delta}$.
  - $p(y \mid \boldsymbol{x}, \boldsymbol{\delta}, \lambda) = \mathcal{N}(y \mid f(\boldsymbol{x} + \boldsymbol{\delta}), \lambda^{-1})$
  - priors (η, a_0, b_0 are hyperparameters):
    - ✓ $p(\boldsymbol{\delta}) = \mathcal{N}(\boldsymbol{\delta} \mid \mathbf{0}, \eta \mathbf{I})$
    - ✓ $p(\lambda) = \mathrm{Gam}(\lambda \mid a_0, b_0)$

- Then, the Bayesian posterior can be viewed as a probabilistic version of LC.
  - Posterior distribution

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



T. Idé, N. 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.

Separating the contribution of each variable needs variational approximation

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

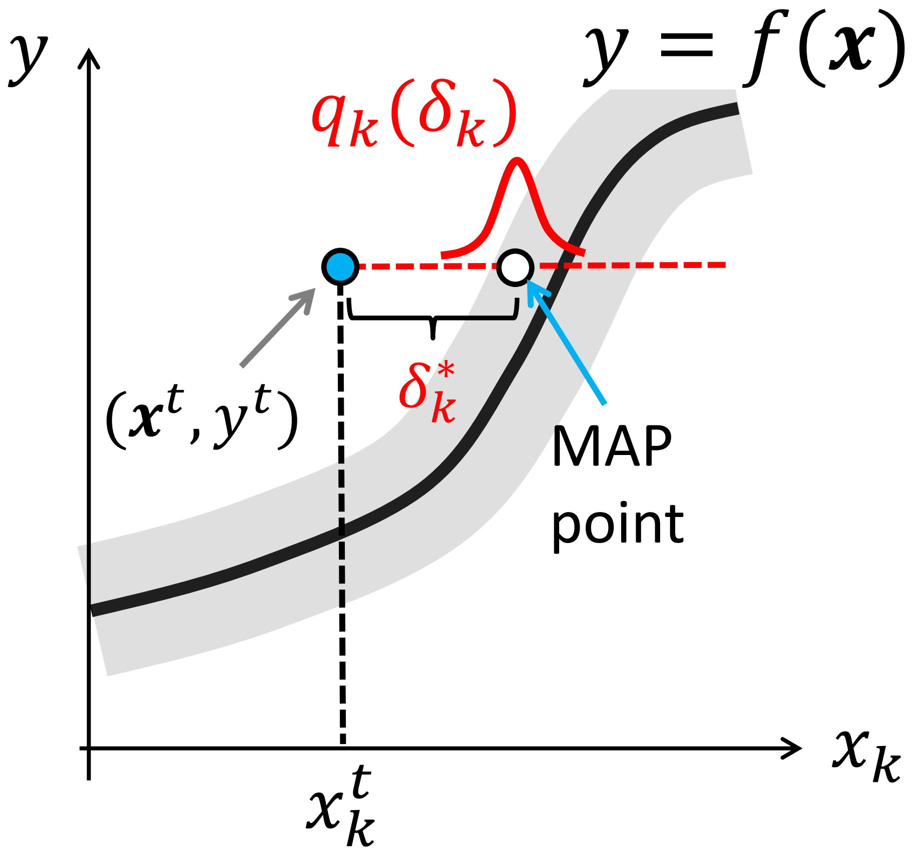
$$Q(\boldsymbol{\delta}) \propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\text{test}}} \int_0^\infty \mathrm{d}\lambda \; p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) p(\lambda),$$

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

- 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 $\boldsymbol{g} = \mathbf{0}$
  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

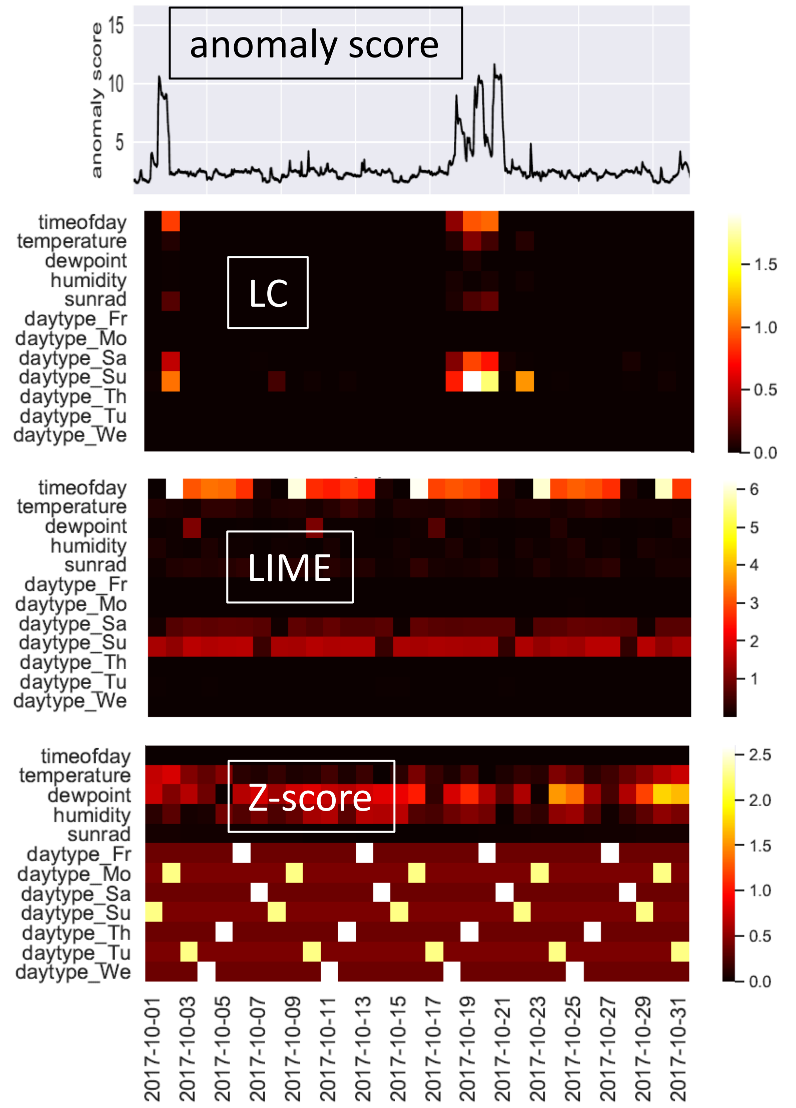
return $\{q_k(\cdot) \mid k = 1, \ldots, M\}$ and $\boldsymbol{\delta}^*$

Agenda

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

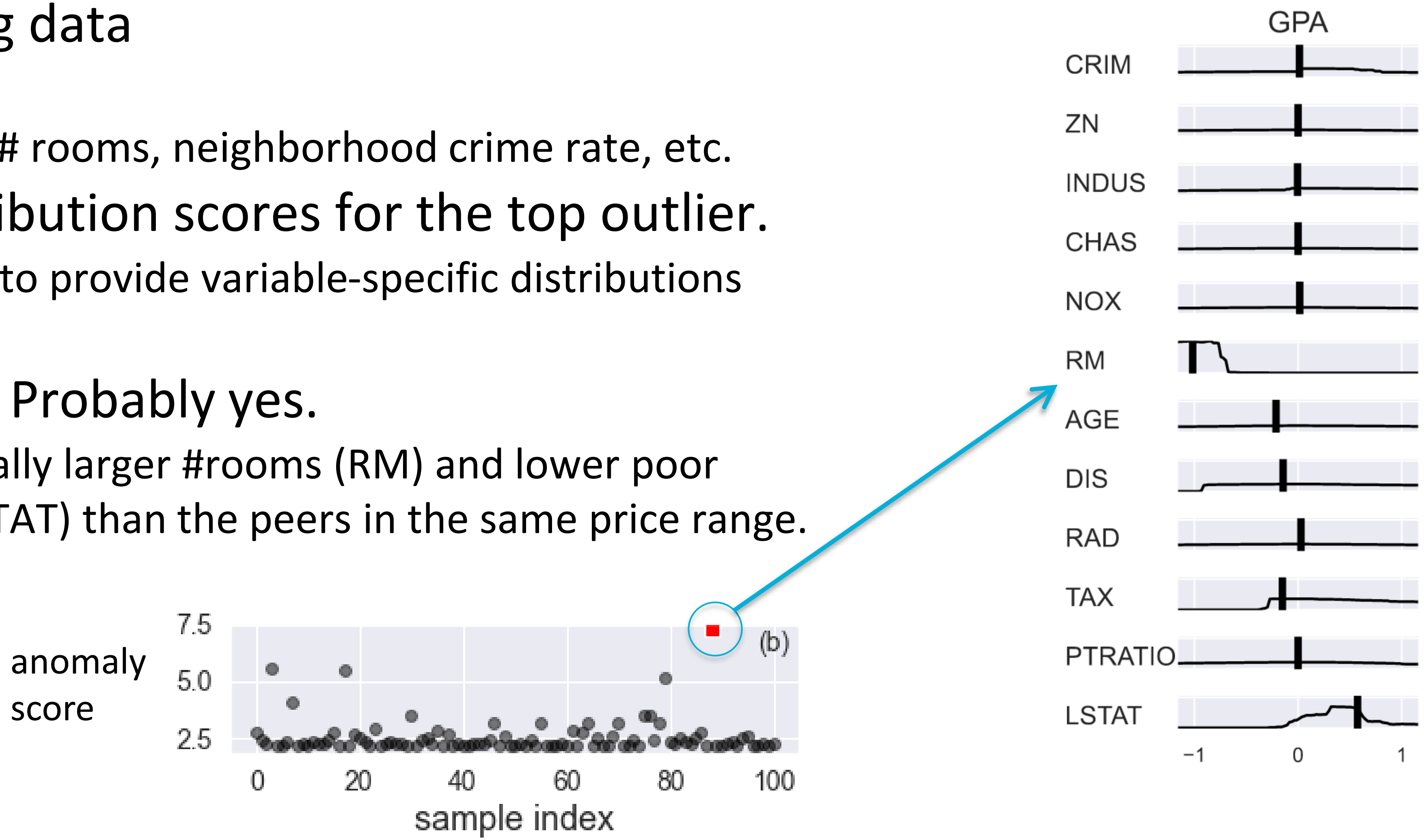
"Why did they exhibit anomalous energy consumption?"
Building energy use-case

- One month-worth building energy data
  - *y*: energy consumption
  - ***x***: time of day, temperature, humidity, sun radiation, day of week (one-hot encoded)
- The score is computed based on hourly 24 test points for each day
  - The mean of the absolute values are visualized
- LC pinpoints the root cause: The big scores on daytime_Su and daytime_Sa imply they look like holidays, which is indeed correct!
- LIME is insensitive to outliers
- Z-score does not depend on *y* (by definition)
  - The artifact for the day-of-week variables is due to one-hot encoding



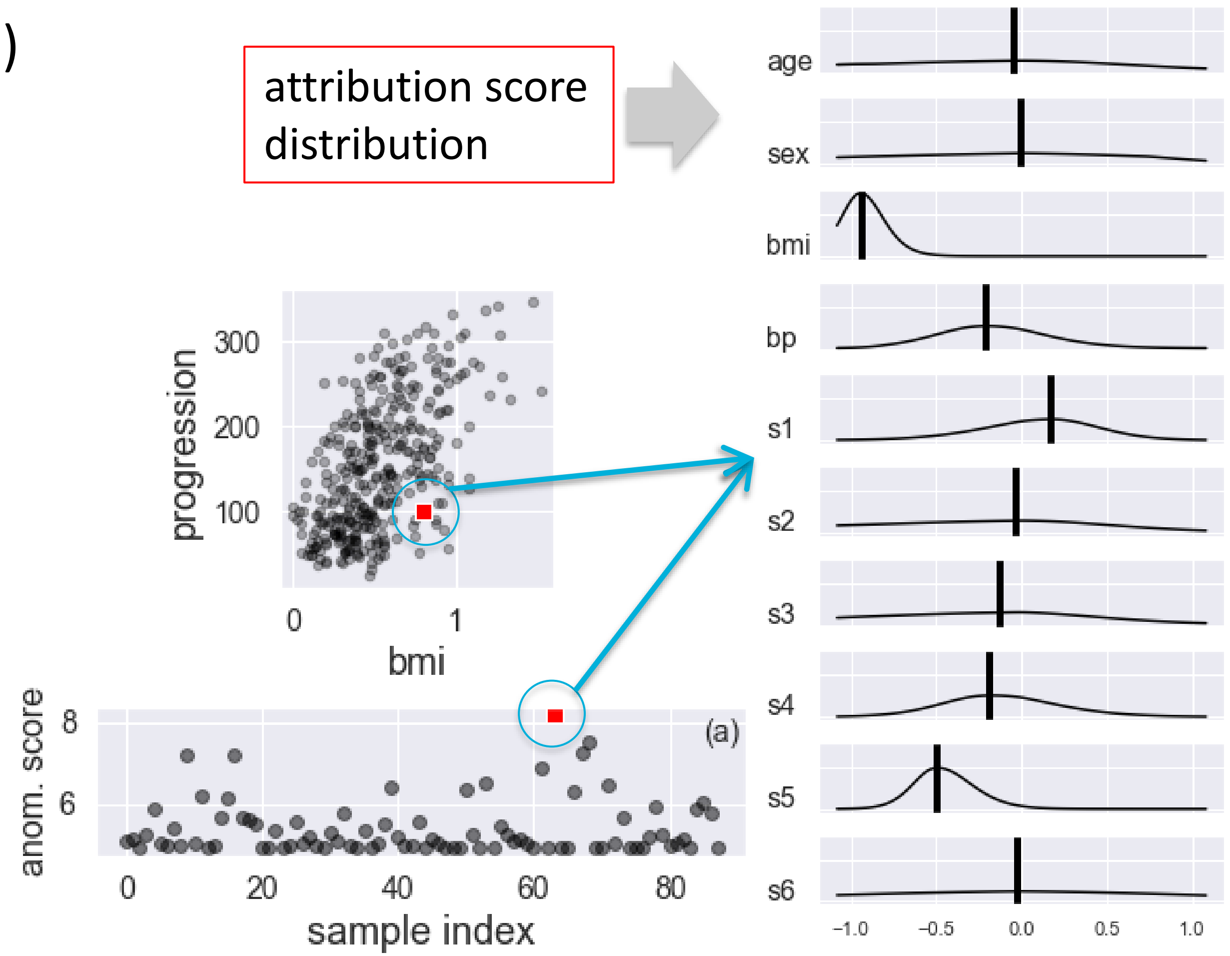
"Why does this house look so unusual?"
House hunting use-case

- Boston Housing data
  - y: house price
  - **x**: house age, # rooms, neighborhood crime rate, etc.
- Computed attribution scores for the top outlier.
  - GPA was able to provide variable-specific distributions
- 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”



Agenda

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

Summary

- Introduced the task of black-box anomaly attribution.
- Rather surprisingly, existing major black-box attribution methods are not capable of explaining deviations.
- Introduced the new notion of likelihood compensation (LC, [Ide+ 21]) and its probabilistic extension (GPA, [Ide-Abe 23]).

Thank you!