IBM **Semiconductors**

Computing Input Responsibility Scores in Black-Box Anomaly Detection

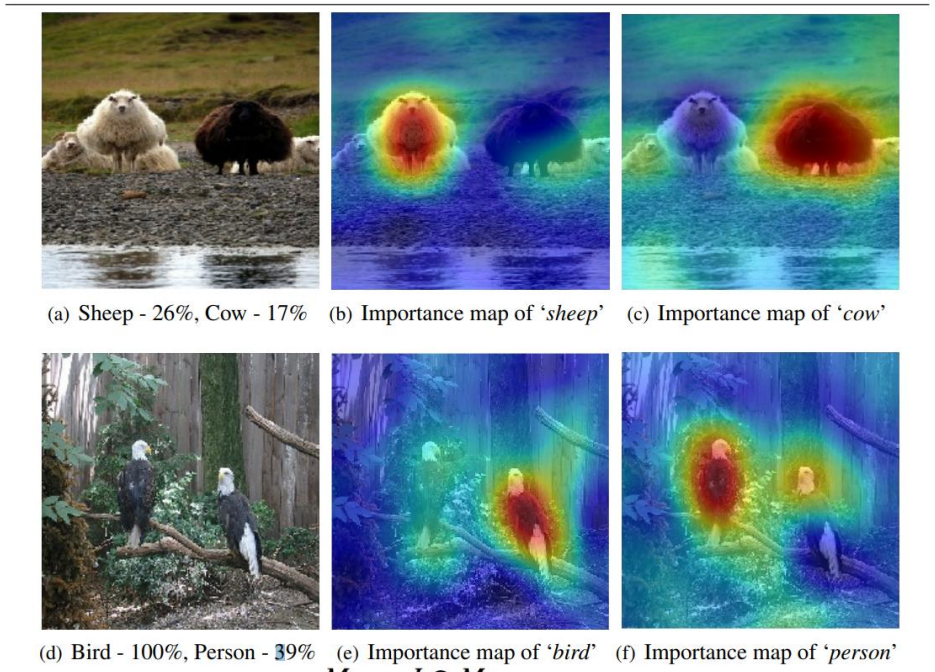
Tsuyoshi (Ide-san) Ide (井手 剛), Ph.D.,
Head of Data Science, IBM Semiconductors, at IBM Research Division

Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

XAI technique: An example (the RISE algorithm)

- Explains why specific object categories are relevant to an input image by showing category-specific saliency map
  - Input image tensor: $\mathbf{X}$
  - Use random binary mask $\mathbf{M}_1, \ldots, \mathbf{M}_N$
  - Saliency map:
    - ✓ $\mathbf{S}^{\text{sheep}} = \frac{1}{N}\sum_{i=1}^{N} p^{\text{sheep}} \cdot \mathbf{M}_i$
  - Sheep probability:
    - ✓ $p^{\text{sheep}} = f^{\text{sheep}}(\mathbf{X} \odot \mathbf{M}_i)$
- Applicable to any black-box image classifier $f^{\text{sheep}}(\cdot), f^{\text{cow}}(\cdot), f^{\text{bird}}(\cdot), \ldots$



Petsiuk, V. "Rise: Randomized Input Sampling for Explanation of black-box models." arXiv preprint arXiv:1806.07421 (2018).

XAI technique: An example (the RISE algorithm)

- Classifier-only
- Explanation = input attribution
  - Which input dimension contributed to the outcome the most
- Model-agnostic
  - Use only the classifier API: $p^k = f^k(\text{image})$
    - ✓ k: The index of a predefined category
    - ✓ p^k: Probability of image belonging to the k-th category
- Instance-specific, i.e., local
  - Don't care about general properties of $f^k(\cdot)$ over the entire input domain.
  - Only properties relevant to a specific image matter.
- No critical hyperparameter

Taxonomy of XAI methods: Complex and multi-faceted without context

- Inherently explainable models
  - Examples: Decision trees, linear models
- Back vs. white box
  - White-box:
    - ✓ Access to the model parameters and / or training samples
  - Black-box:
    - ✓ Only prediction outcome
    - ✓ Often not have access to training samples
- Model-agnostic vs. model-specific
  - Model-specific:
    - ✓ Uses intermediate outputs (e.g., feature maps, embeddings)
  - Model-agnostic (post-hoc):
    - ✓ Can be applied to any model, regardless of internal structure
- Local vs. global explanation
  - Local:
    - ✓ For a specific input sample
  - Global:
    - ✓ Explains model behavior in a subdomain capture characteristics of the model in a subdomain
      - e.g., using a Gaussian mixture model that is trained in that subdomain.
- Surrogate model
  - A simpler, easy-to-explain model fitted to approximate a more complex model in a subdomain
- Feature attribution vs. example-based
  - Feature attribution:
    - ✓ Explains prediction based on feature contributions
    - ✓ Example-based:
      - "Prediction of X was Y because X's features are similar to X1, which also belongs to Y"

Speith, Timo. "A review of taxonomies of explainable artificial intelligence (XAI) methods." *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2022.

Explanation is context-dependent. Ability to provide actionable insight is the key.

- Question: What is a good explanation?
- My answer (as a professional industry researcher):
  - It entirely depends on the downstream business process.
    - ✓ An explanation is useful if it provides clearer, actionable insights.
    - ✓ There is no universally good explanation.
  - Subjective satisfaction from general public is not relevant, in my opinion.

- Different Perspectives:
  - Some argue that user studies are the only way to assess the quality of explanations.
  - Others believe that reducing explanations to a well-known benchmark is essential.

Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

Focusing on input attribution for local, black-box *regression* model

- Most of XAI methods focus on the classification task
  - particularly for images.
  - Many industrial applications don't fit neatly into classification problems.

- In regression models, especially when training samples are unavailable, **input attribution** can be viewed as the primary approach for explanations.

Example: **Wine quality prediction**

Why is this particular wine rated as the highest quality?

Which predictor variables (x) contribute to the outcome variable (y) most?

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

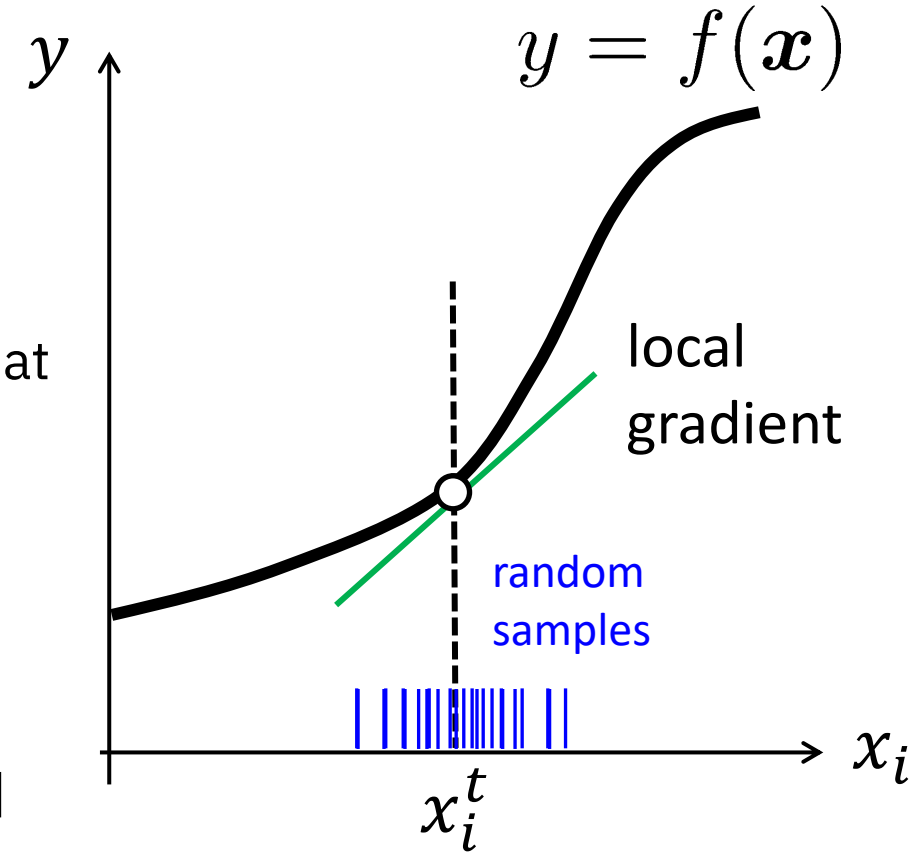
$$y = f(\boldsymbol{x})$$

Reviewing three existing methods and their variants

- Local linear surrogate modeling (aka LIME)
- Integrated gradient (IG)
  - Expected integrated gradient (EIG)
- Shapley values (SV)
  - Kernel SHAP

LIME performs local sensitivity analysis on black-box functions

- Sensitivity = gradient = attribution score
- Challenge:
  - $f(\boldsymbol{x})$ is black-box; No way of computing the gradient analytically.
- Idea:
  - Randomly generate samples around a test sample $\boldsymbol{x}^t$ at which you want to obtain a model explanation.
    - ✓ $\{(\boldsymbol{x}^{t[1]}, y^{t[1]}), \ldots, (\boldsymbol{x}^{t[N]}, y^{t[N]})\}$ 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 serve an estimator of the gradient (= explanation).
- The regression plane can be viewed as a simplified surrogate model defined locally.



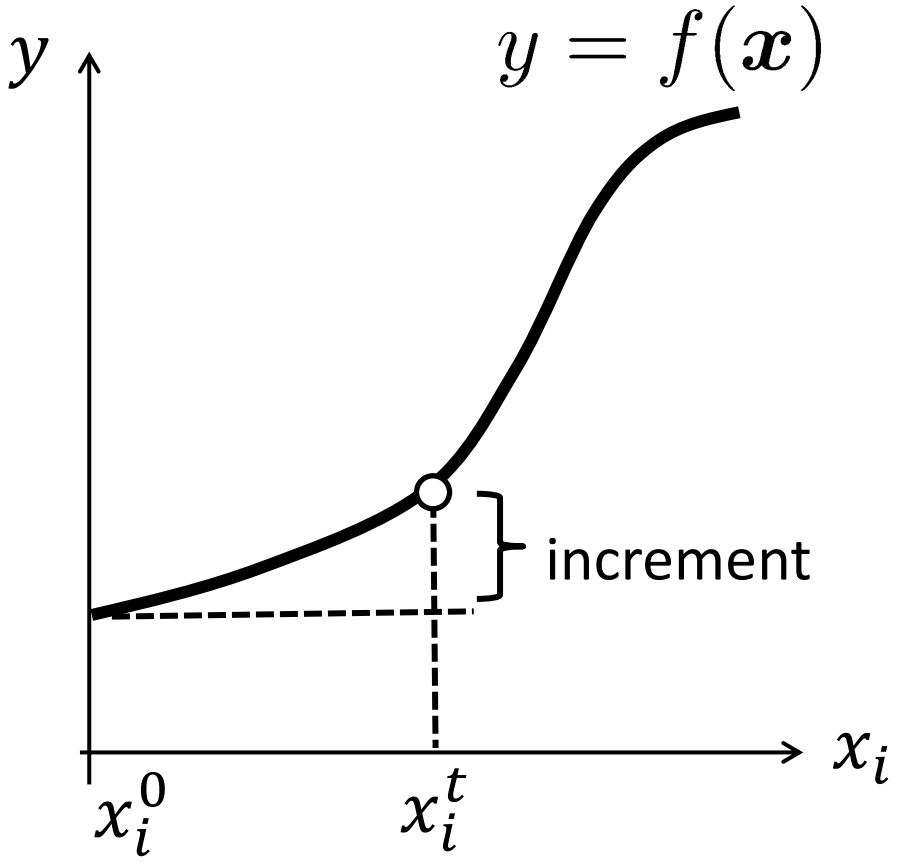
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).
- LIME: Local Interpretable Model-agnostic Explanations

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

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

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

    - ✓ Gradually changing α from 0 to 1 amounts to a shift from the "reference point" to the test sample of interest $\boldsymbol{x}^t$
    - ✓ Integration = collecting infinitesimal increments
- This is another local attribution method.
- Issue: dependency on the arbitrary ref point.



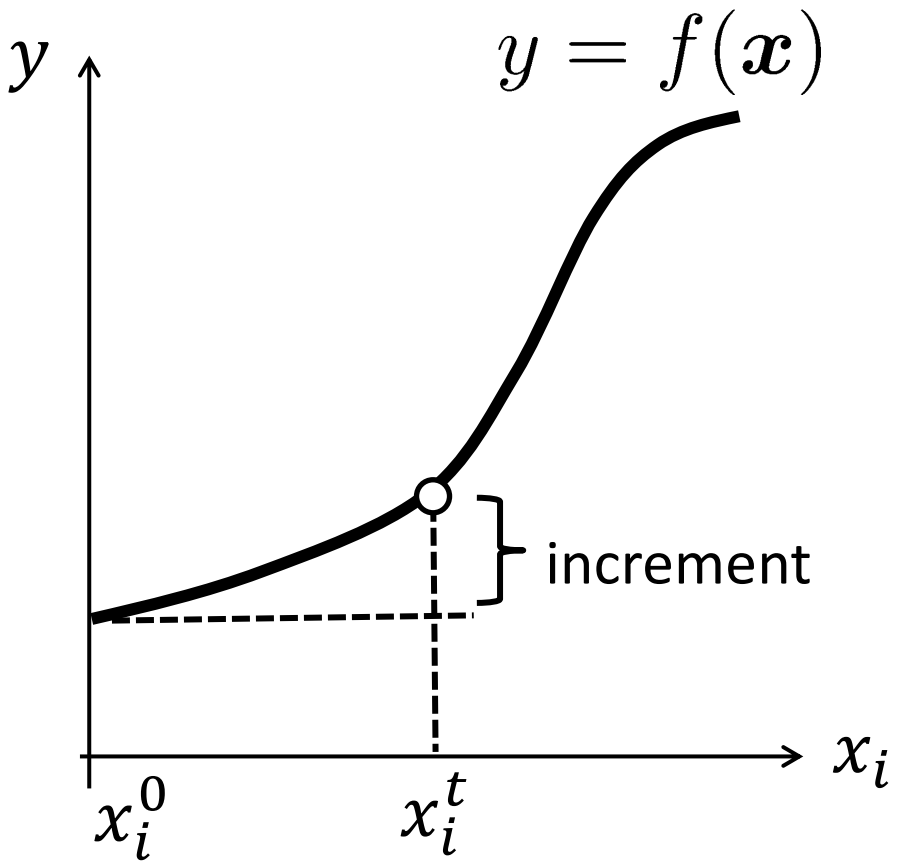
- John Sipple. "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure," In Proceedings of the 37th International Conference on Machine Learning (ICML 20).

Expected IG (EIG) eliminates the dependency on the arbitrary reference point

- Expected IG (EIG) [Deng+ 21]
  - Computed by marginalizing $\boldsymbol{x}^0$ with a distribution of the reference point

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

    - The empirical distribution of the training samples, which is often unavailable.
    - LIME-like sampling distribution leads to a meaningless EIG value due to mutual cancelling.
- Increments and gradients are closely related. Can we relate (E)IG with LIME mathematically?
  - Yes → shown later



- Huiqi Deng, et al. , A Unified Taylor Framework for Revisiting Attribution Methods. In Proceedings of the AAAI Conference on Artificial Intelligence. 11462–11469, 2021.

Shapley values (SV): Evaluating the impact of each variable's "participation"

- General definition of the SV (of the i-th input variable)

$$\mathrm{SV}_i(\boldsymbol{x}^t) = \sum_{S^{(i)}} \mu(\mathcal{S}^{(i)}) \left[v(\mathcal{S}^{(i)}) - v(\mathcal{S}^{(i)} - i) \right]$$

  - $\mathcal{S}^{(i)}$: A set of variables including the i-th
  - $\mathcal{S}^{(i)} - i$: A set of variables removing the i-th from $\mathcal{S}^{(i)}$
  - $v(\cdot)$: The gain function (or characteristic function) quantifying the benefit of forming the specified coalition.
  - $\mu(\mathcal{S}^{(i)})$: importance weight of the configuration $\mathcal{S}^{(i)}$.
- $\mathcal{S}^{(i)}$ specifies a team of coalition.
- Typical choice for $v(\cdot)$ and $\mu\left(\mathcal{S}^{(i)}\right)$
  - $v(\cdot)$ is chosen as the regression function $f(\cdot)$ itself [Štrumbelj- Kononenko 14].
  - $\mu\left(\mathcal{S}^{(i)}\right) = \left(M \times \binom{M-1}{k}\right)^{-1}$ with $k = |\mathcal{S}^{(i)}|$ and M is the number of input variables.

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

Shapley values (SV): Understanding the notion of "participation"

- Typical definition:
  - Participation = take the value of $\boldsymbol{x}^t$ (test sample of interest)
  - Non-participation = take the value of some reference point
- 4-variate regression function example (i.e., M=4)
  - For $i = 1, k = 3$, and $\mathcal{S}^{(1)} = \{1,2,3\}$,
    - ✓ $\Delta f(\mathcal{S}^{(i)}) \equiv f(\mathcal{S}^{(i)}) - f(\mathcal{S}^{(i)} - i) = f\left(x_1^t, x_2^t, x_3^t, x_4^{(n)}\right) - f\left(x_1^{(n)}, x_2^t, x_3^t, x_4^{(n)}\right)$
    - ✓ Here, $\boldsymbol{x}^{(n)}$ is a reference point
  - Typically, the reference point is integrated out using the empirical distribution:
    - ✓ $\Delta f(\mathcal{S}^{(i)}) = \frac{1}{N}\sum_{n=1}^{N}[f\left(x_1^t, x_2^t, x_3^t, x_4^{(n)}\right) - f\left(x_1^{(n)}, x_2^t, x_3^t, x_4^{(n)}\right)]$
- SV under the standard definition

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

Shapley values (SV): Handling computational challenges

- SV under the standard definition

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

- When *M* (input dimensionality) is large, exact computation is prohibitively expensive.
- Typical approximation methods
  - Monte Carlo evaluation [Štrumbelj- Kononenko 14].
  - kernelSHAP [Lundberg-Lee 17]
    - ✓ Leverages the (fascinating) characterization of SV as the solution of a least squares problem (!) [Charnes+ 88].
    - ✓ (Details omitted)

- Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017).
- Charnes, A., et al. "Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations." Econometrics of planning and efficiency (1988): 123-133.

Unifying LIME, (E)IG, and SV into the same family [Ide+ 23]

- EIG and SV satisfy the same sum rule:

$$\sum_{i=1}^{M} \mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0) = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^0), \quad \sum_{i=1}^{M} \mathrm{EIG}_i(\boldsymbol{x}^t) = \sum_{i=1}^{M} \mathrm{SV}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - \langle f \rangle$$

$\langle f \rangle$: $f(\cdot)$'s average in the domain

  - ✓ $\boldsymbol{x}^t$: test point (at which you want to get an explanation)
  - ✓ $\boldsymbol{x}^0$: reference point
- EIG and SV are equivalent up to the second order of Taylor expansion.
  - $\mathrm{EIG}_i(\boldsymbol{x}^t) \approx \mathrm{SV}_i(\boldsymbol{x}^t)$
- LIME can be viewed as the gradient of EIG and IG in a certain limit

$$\mathrm{LIME}_i(\boldsymbol{x}^t) = \frac{\partial \mathrm{EIG}_i(\boldsymbol{x}^t)}{\partial x_i} = \lim_{\boldsymbol{x}^0 \to \boldsymbol{x}^t} \frac{\partial \mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0)}{\partial x_i},$$

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

Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

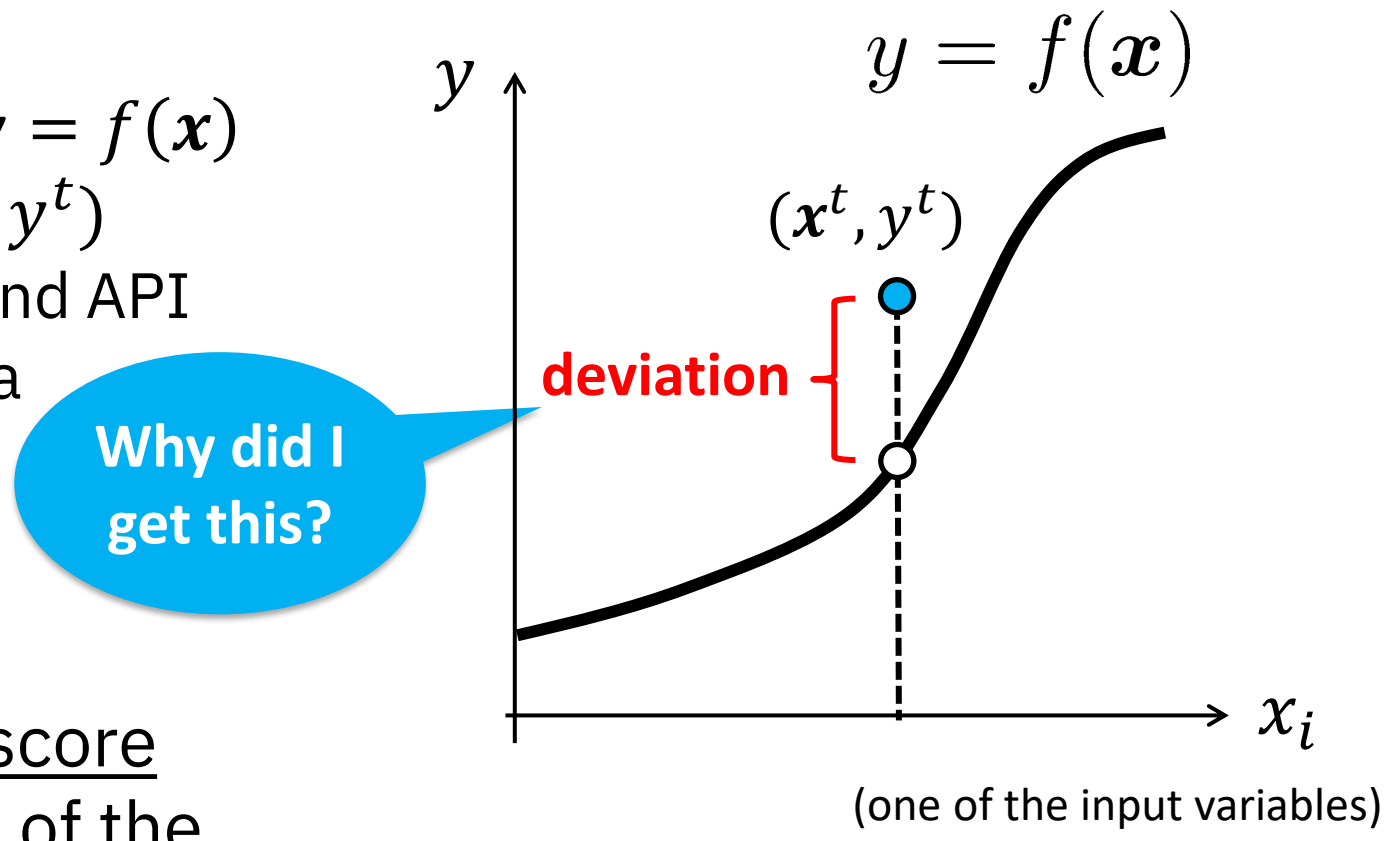
Explaining deviations between prediction and observation

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

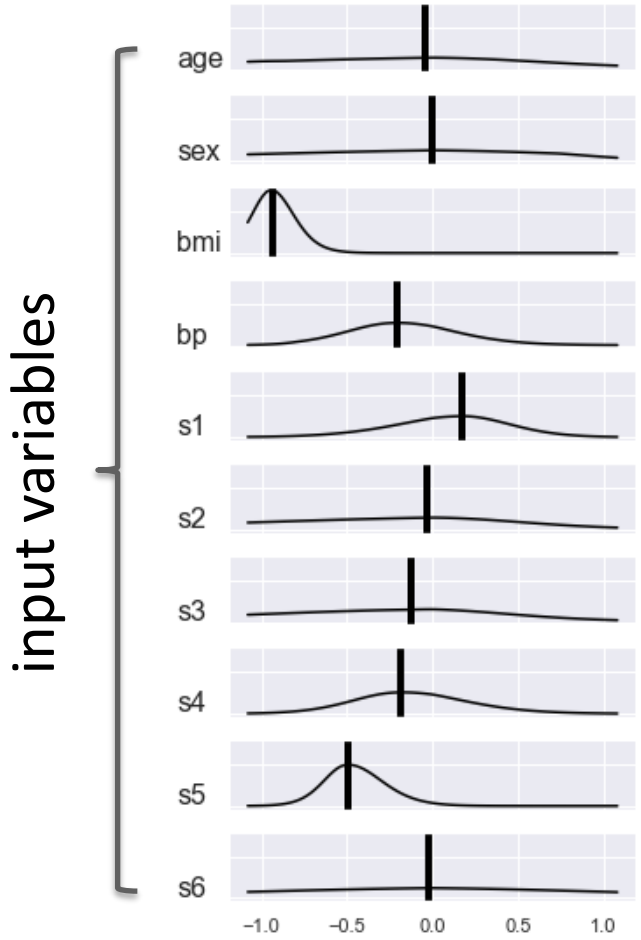
Explain:

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

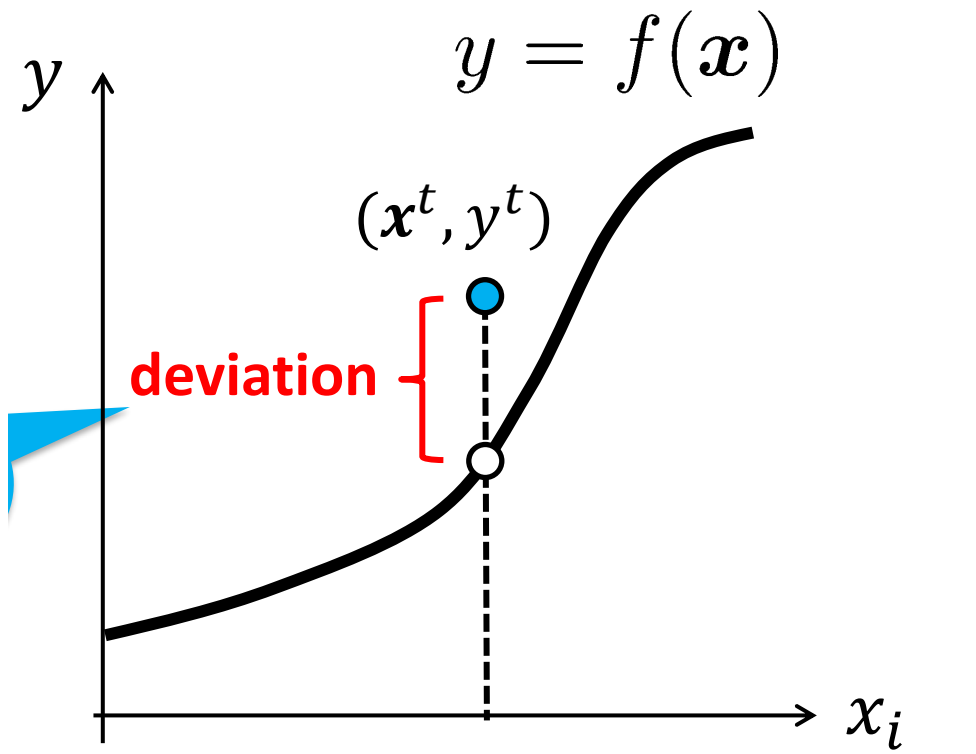


Explaining deviations between prediction and observation

Hopefully, we want to get the score's confidence as well.



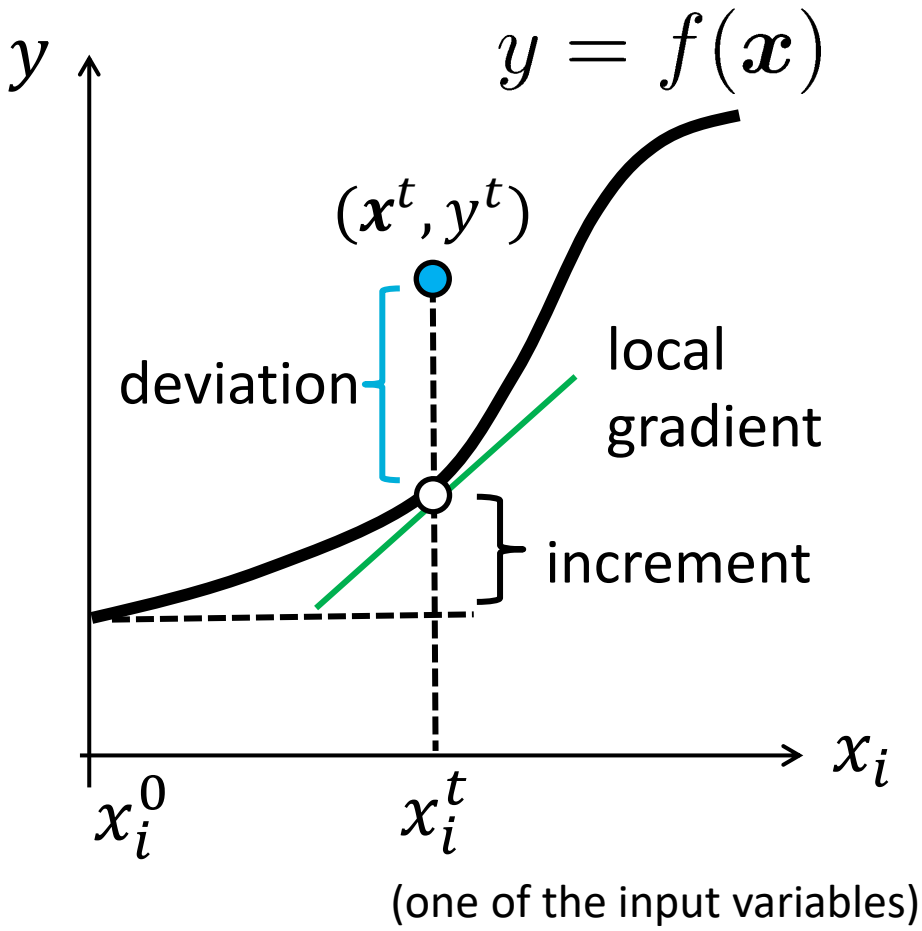
attribution score (responsibility)



(one of the input variables)

Explaining deviations is fundamentally different from explaining the model: Strong impossibility results

- LIME, SV, IG, and EIG are deviation-agnostic [Ide+ 23].
  - These methods explain $f(\boldsymbol{x})$ locally at $\boldsymbol{x} = \boldsymbol{x}^t$, independently of the observed outcome value y.
  - The limitation remains true even if we aim to explain the function $f(\boldsymbol{x}) - y$ rather than just $f(\boldsymbol{x})$.
- Intuition (→ illustration):
  - LIME as local gradient has nothing to do with the deviation.
  - The increment of the regression function is unrelated with the deviation.

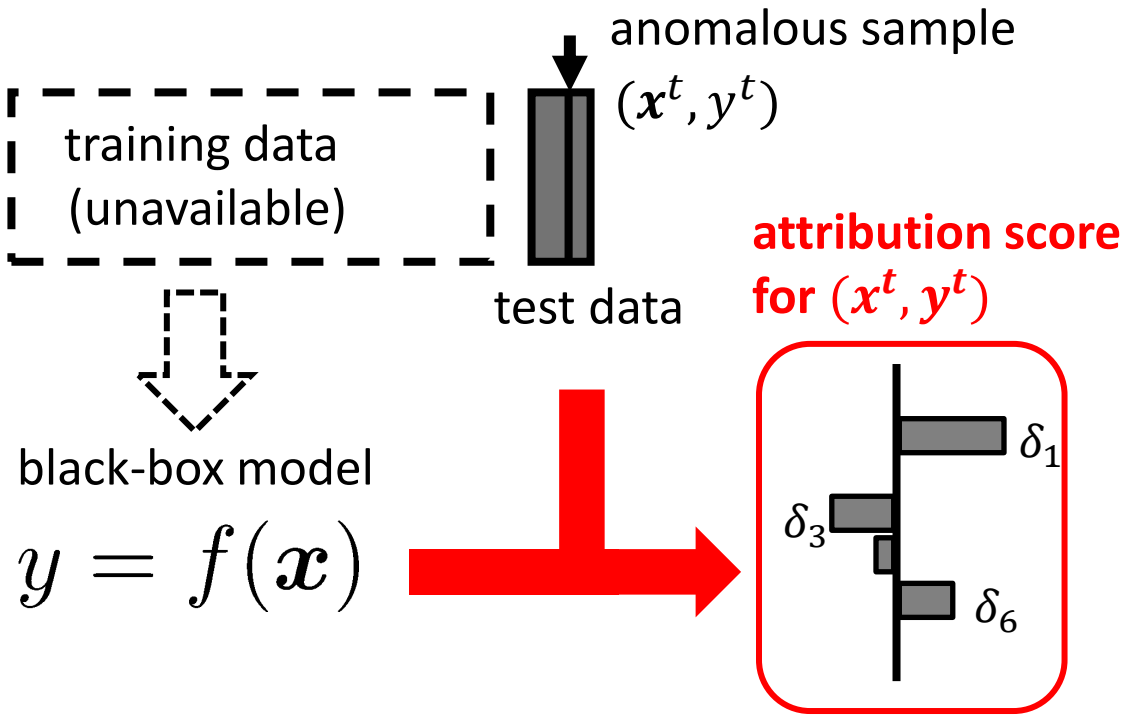


T. Idé, N Abe, "Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution," In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023, August 6-10, 2023, Long Beach, California, USA), pp. 845-856.

Seeking a new paradigm beyond local linear approximation

- Observations so far:
  - LIME, (E)IG, and SV belong to the same family, relying on local linear approximations.
  - These methods are inherently deviation-agnostic.

- What's next:
  - We will briefly review key ideas from [Ide et al. AAAI 21] and [Ide et al. KDD 23] for moving beyond these limitations.

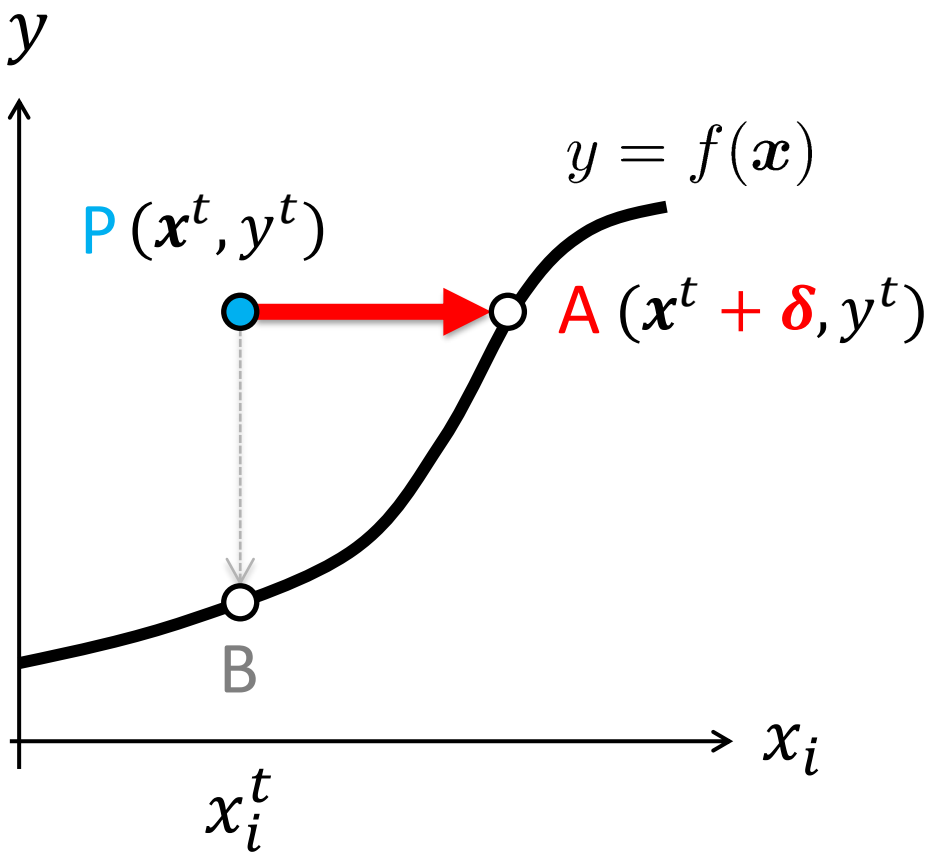


Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

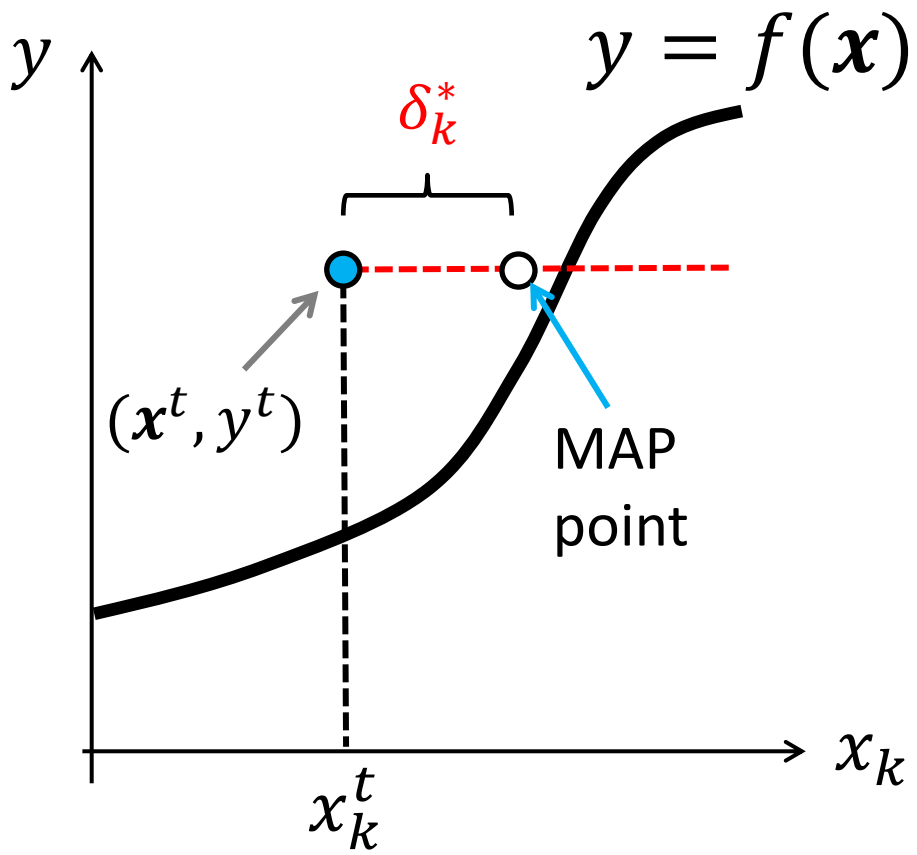
Given a test point $(\boldsymbol{x}^t, y^t)$ being anomalous, we ask: How much "work" would we need to bring it to the normalcy?

- 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, $N_{\mathrm{test}} = 1$



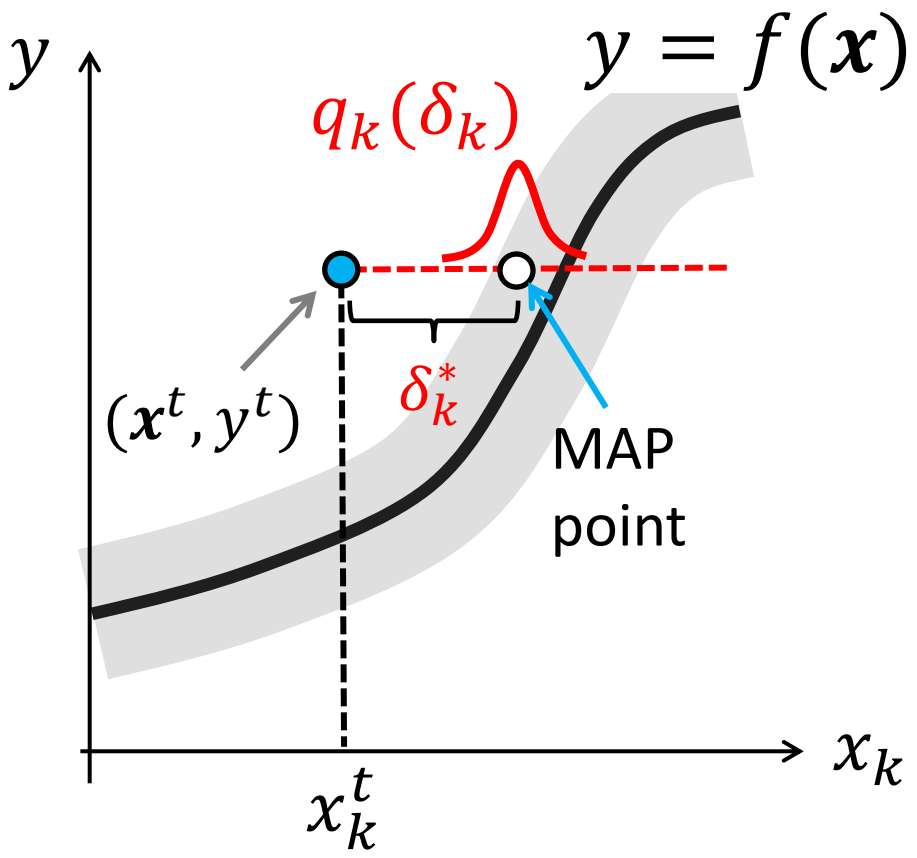
T. Idé, et al., Naoki Abe, "Anomaly Attribution with Likelihood Compensation," In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 21, February 2-9, 2021, virtual), pp.4131-4138

Generative perturbation analysis (GPA) [Ide+ 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 $N_{\text{test}} = 1$)

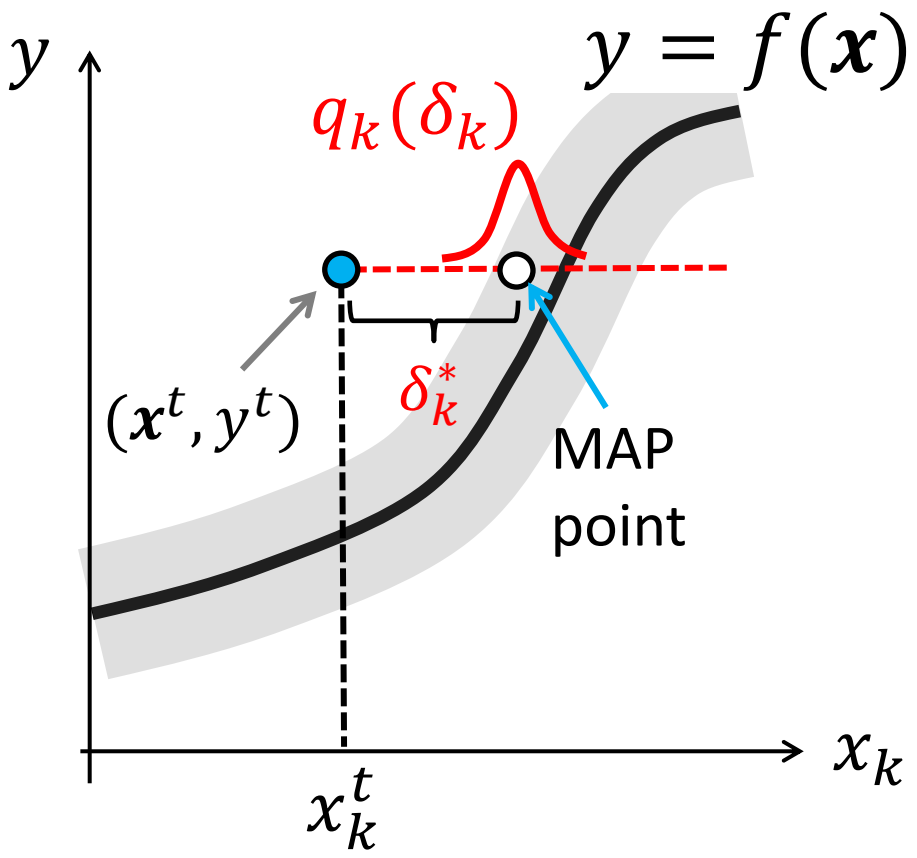
$$Q(\boldsymbol{\delta}) \propto p(\boldsymbol{\delta}) \prod_{t=1}^{N_{\text{test}}} \int_0^\infty \mathrm{d}\lambda \; p(y^t \mid \boldsymbol{x}^t, \boldsymbol{\delta}, \lambda) p(\lambda),$$

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

- 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 $\boldsymbol{\delta} \approx \mathbf{0}$.

repeat — MAP

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

for all k **do** — distribution

  $q_k(\delta) = Q(\delta_1^*, \ldots, \delta_{k-1}^*, \delta, \delta_{k+1}^*, \delta, \delta_M^*)$

  $q_k(\cdot) \leftarrow q_k(\cdot) / \int \mathrm{d}\delta' q_k(\delta')$ with Eq. (18)

end for

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

Comparing GPA with other methods: Summary

- LC and GPA associate the likelihood function with normalcy.
  - Assumption: likely = normal
- This makes the algorithm deviation-sensitive and eliminates the dependency on arbitrary reference points.

Table 1: Comparison of model-agnostic attribution methods in the regression setting.

	model-agnostic	training-data-free	baseline-input-free	y-sensitive	built-in UQ	reference point
LIME [33]	yes	yes	yes	no	yes/no	infinitesimal vicinity
SV [41, 42]	yes	no	yes	no	no	globally distributional
IG [37, 44]	yes	yes	no	no	no	arbitrary
EIG [6]	yes	no	yes	no	no	globally distributional
Z-score [5]	yes	no	yes	no	no	global mean of predictors
LC [20]	yes	yes	yes	yes	no	maximum likelihood point
GPA	**yes**	**yes**	**yes**	**yes**	**yes**	maximum a posteriori point

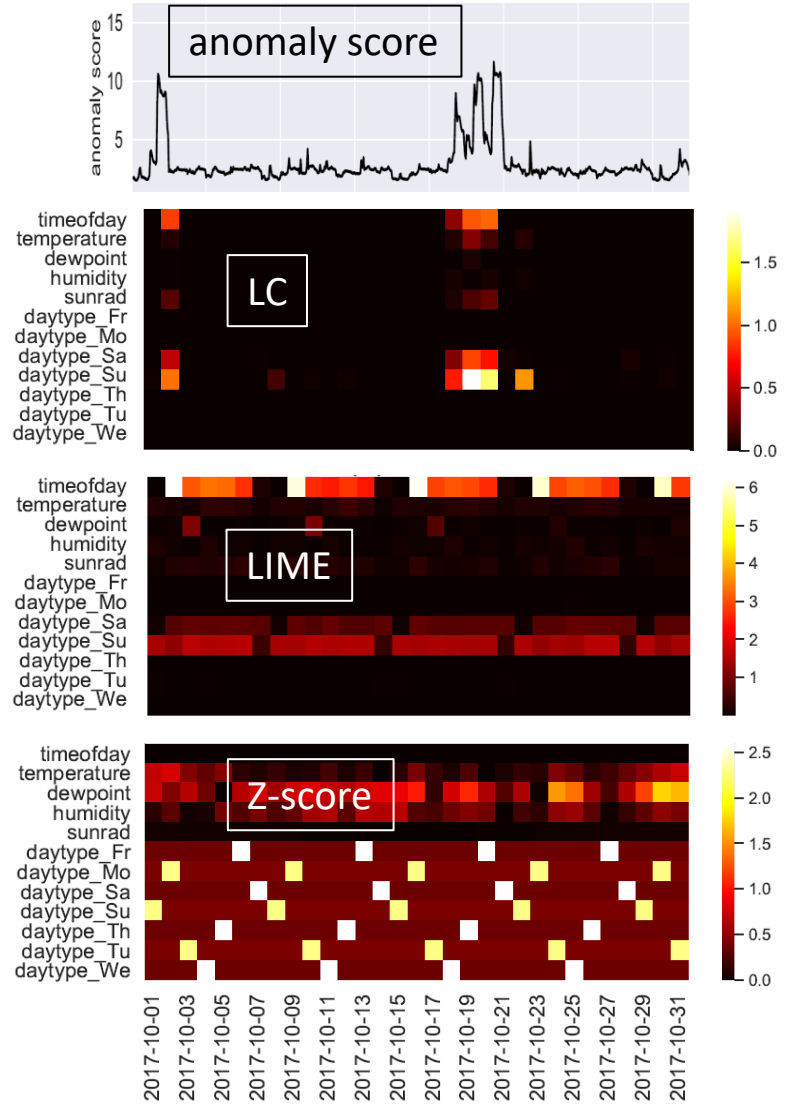
Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

“Why did my building’s AC system look anomalous?” Building energy use-case

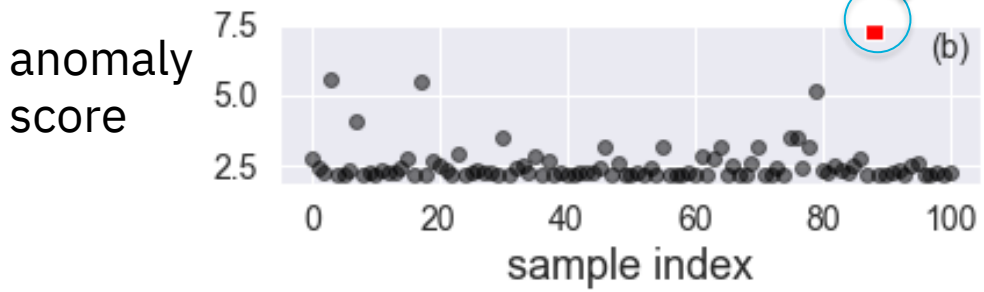
TRIRIGA® An IBM Company

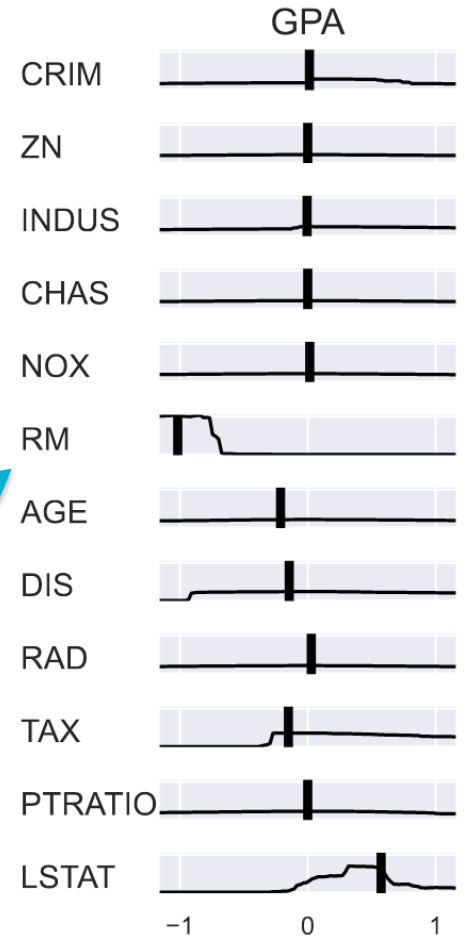
- One month-worth building energy data
  - *y*: energy consumption
  - ***x***: time of day, temperature, humidity, sun radiation, day of week (one-hot encoded)
- The score is computed based on hourly 24 test points for each day
  - The mean of the absolute values are visualized
- LC pinpoints the root cause: High scores for daytime_Su (Sunday) and daytime_Sa (Saturday) suggest these days behave like holidays, which is accurate.
- 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

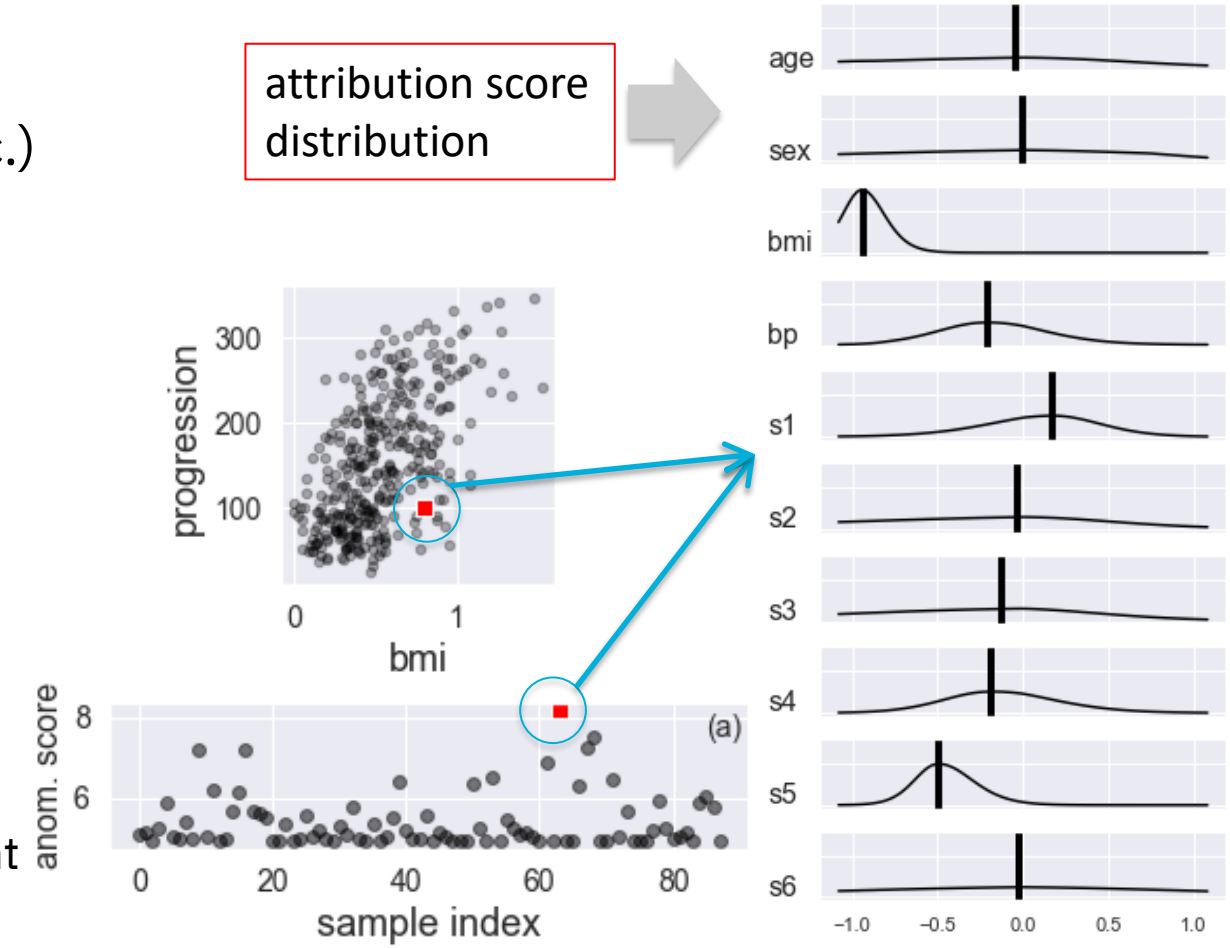
- Boston Housing data
  - y: house price
  - **x**: house age, # rooms, neighborhood crime rate, etc.
- Computed attribution scores for the top outlier.
  - GPA was able to provide variable-specific distributions
- Is it a bargain? Probably yes.
  - The house has unusually more rooms (RM) and a lower percentage of poor neighbors (LSTAT) compared to others 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, the patient would appear typical if his BMI were much lower.
  - He could be described as:
    - ✓ Overweight but with low progression
    - ✓ Low progression despite being overweight



Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
  - Generative perturbation analysis (KDD 23)
  - Results examples
- Application to advanced semiconductor manufacturing

Wafer defect density prediction is a huge black-box regression task

- Wafer defect density prediction as a function of process parameters (waiting time, recipe, etc.) and measurements (electronic resistance, etc.).
  - Mfg. process is so complex that precise physics modeling is not possible.
  - Data-driven models (e.g., DNN) should play a critical role for prediction.

- Deviation explanation can be viewed as the inverse problem of the regression task.

AI/ML themes we are working on

Simulation-based fab optimization

Fleet-level modeling of tool's sensor data

Trajectory-based process diagnosis

Thank you!

Computing Input Responsibility Scores in Black-Box Anomaly Detection

- Abstract:
  - Explainable AI (XAI) is an active research field in machine learning (ML) aimed at addressing growing concerns about the black-box nature of deep learning models. One particularly interesting scenario in XAI is explaining what might go wrong with the model when a prediction significantly deviates from the actual outcome. While this problem can be formalized in many different ways, I focus on the task of input attribution, which seeks to quantify how much each input variable of the model is responsible for the observed deviation.
  - In this talk, I will first review existing attribution approaches recently developed in the ML community, including linear surrogate modeling, Shapley values, and integrated gradients. After summarizing the challenges of these methods in the context of anomaly attribution, I will introduce a newer notion of likelihood compensation as a major counterfactual-type explanation, along with its probabilistic extensions. If time permits, I will also share challenging real-world anomaly attribution problems from semiconductor manufacturing at IBM.
- Bio:
  - Dr. Tsuyoshi ("Ide-san") Ide is currently the head of data science for IBM Semiconductors at IBM Research. He received his Ph.D. in theoretical physics in 2000 from the University of Tokyo, Japan. After joining IBM Research – Tokyo, he shifted his research focus to data mining and machine learning. In 2013, he transferred to the T.J. Watson Research Center in New York. Dr. Ide is passionate about modeling real-world business problems using advanced machine learning techniques and has led numerous customer engagements. His recent research interests include anomaly detection and explanation, point process modeling of discrete events, and analytics of graph-structured data.