IBM **Semiconductors**

Explaining what went wrong: The black-box anomaly attribution problem

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

IEOR Seminar, Department of Industrial Engineering and Operations Research, Columbia University (November 15, 2024).

My background and current role

- Ph.D. in physics (U. Tokyo, Japan) in 2000
  - Theoretical/numerical study on high critical temperature superconductivity
- IBM Research – Tokyo
  - LCD display → machine learning (ML)
- IBM T. J. Watson Research Center (2013-)
  - Applied & basic research of ML
- IBM Semiconductors (2023-)
  - Head of Data Science

- Recent work (https://ide-research.net/publications/)
  - XAI for anomaly detection (AAAI 19, AAAI 21, KDD 23)
  - Point processes (NeurIPS 21, AISTATS 24)
  - Graph neural networks (AAAI 22, ICASSP 23)
  - Decentralized learning (IJCAI 19, SMDS 21)

Strategy areas of IBM Research

Semiconductors | **Artificial Intelligence** | **Quantum Computing** | **Hybrid Cloud**

Logic Technology

Designing the next generation of chips to increase performance and improve energy efficiency.

Chiplet & Adv Packaging

Developing chiplet and packaging architectures built for next generation AI.

Design & Enablement on cloud

Enabling partners with our expertise across chip design and hybrid cloud.

Intelligent Fab

Using AI and automation to make semiconductor manufacturing faster and smarter.

Data science problems in semiconductor manufacturing (preview)

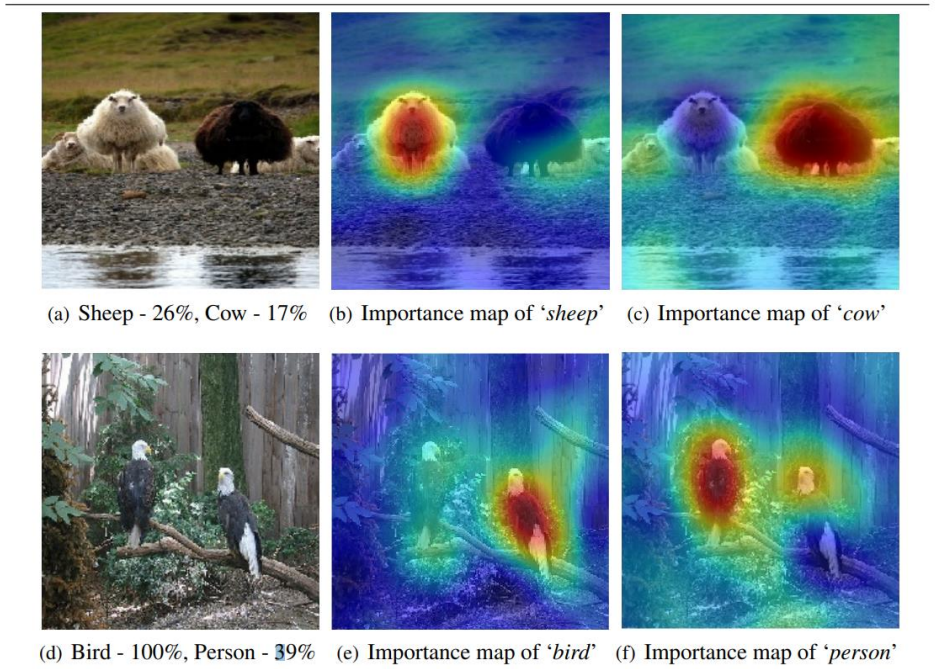
deleted

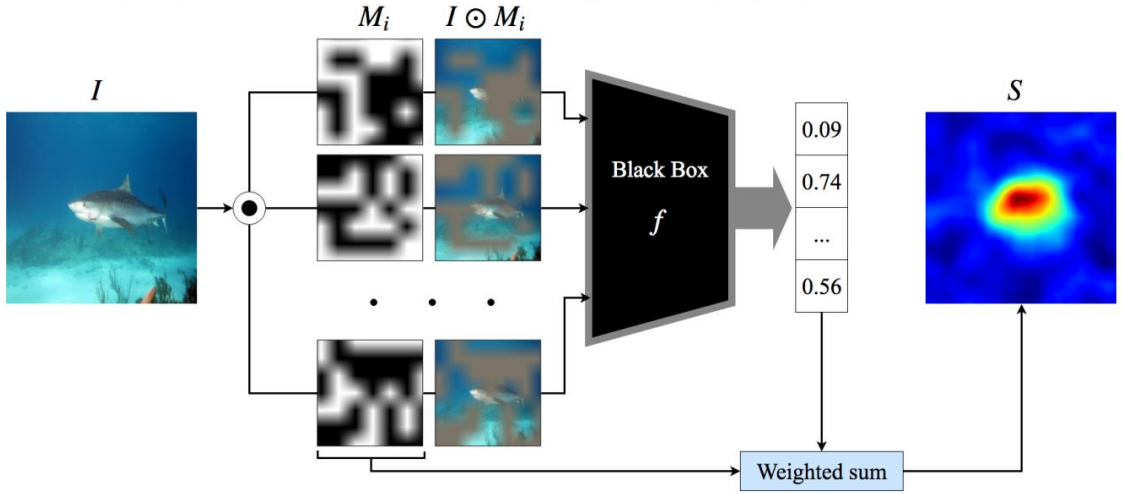
Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
- Data science problems in semiconductor manufacturing

The attribution problem: An example (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).

The attribution problem: An example (RISE algorithm)

- Explanation is provided as input attribution in this case.
  - Which input dimension contributed to the outcome the most

- RISE is one of many possible way of attribution.
  - Valid only for classification.
  - 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 (instance-level):
    - ✓ For a specific input sample
    - ✓ c.f. population-level explanation
  - Global:
    - ✓ Explains model behavior in a 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 (subjective satisfaction) are the only way to assess the quality of explanations.
  - Others believe that reducing explanations to a well-known performance metric (such as classification accuracy) 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 business and 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

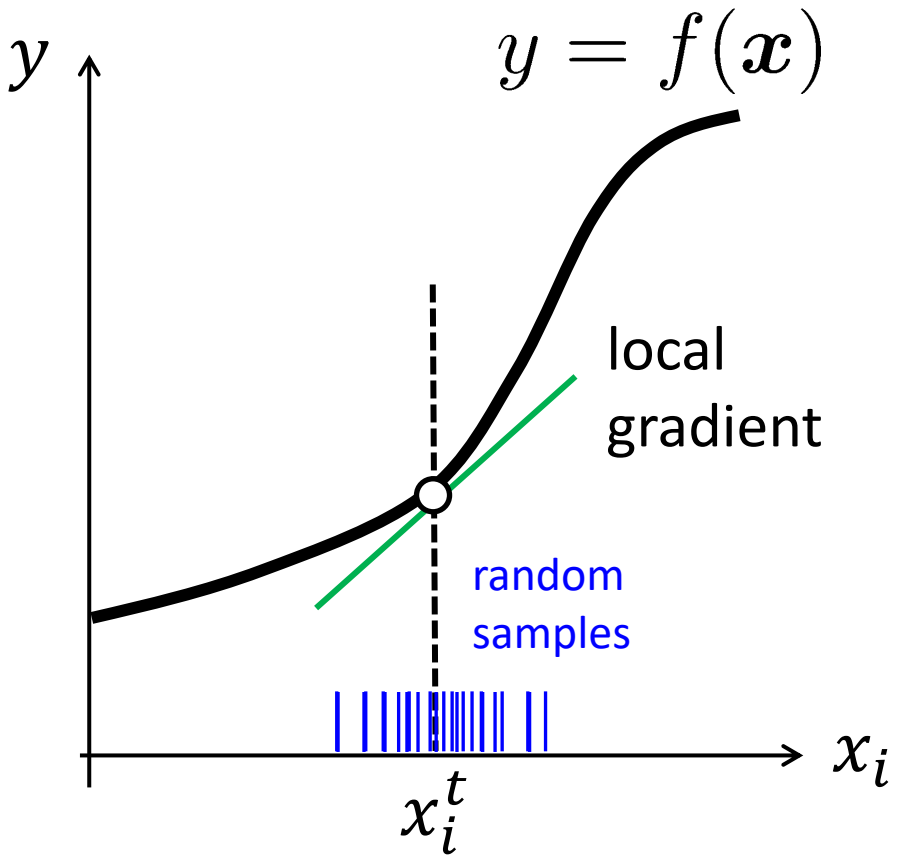
$$y = f(\boldsymbol{x})$$

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 computes the gradient of black-box functions locally.

- Sensitivity = gradient = attribution score
- Challenge:
  - $f(\boldsymbol{x})$ is black-box (i.e., only its API is available). Impossible to compute 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.



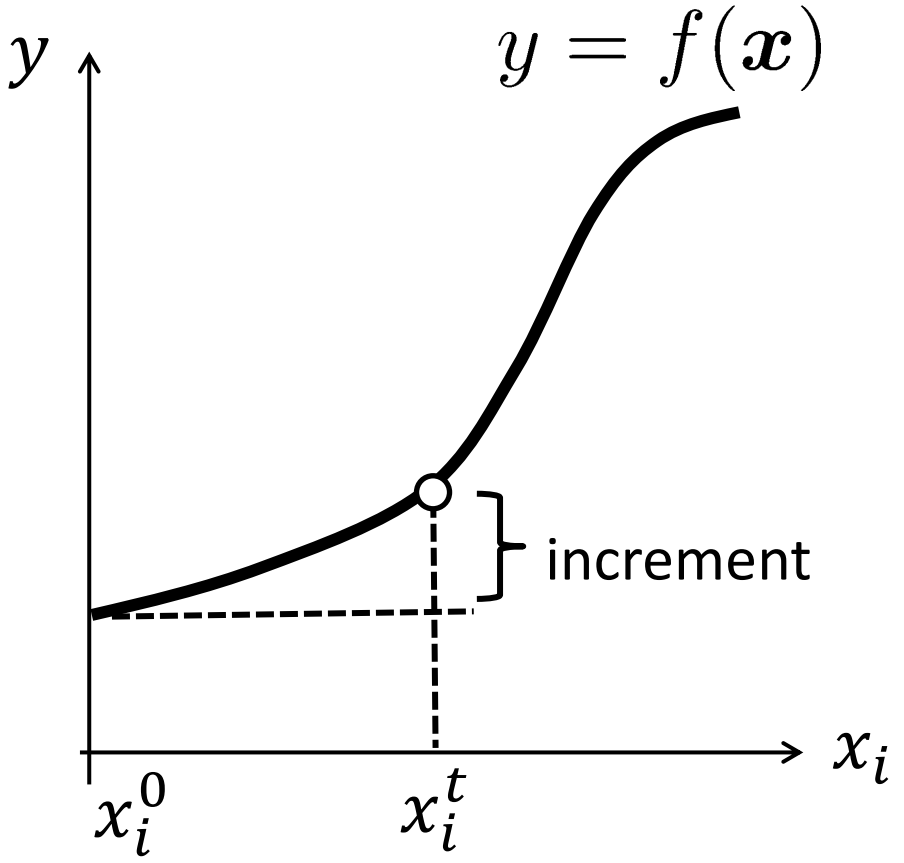
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).
- LIME: Local Interpretable Model-agnostic Explanations

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.



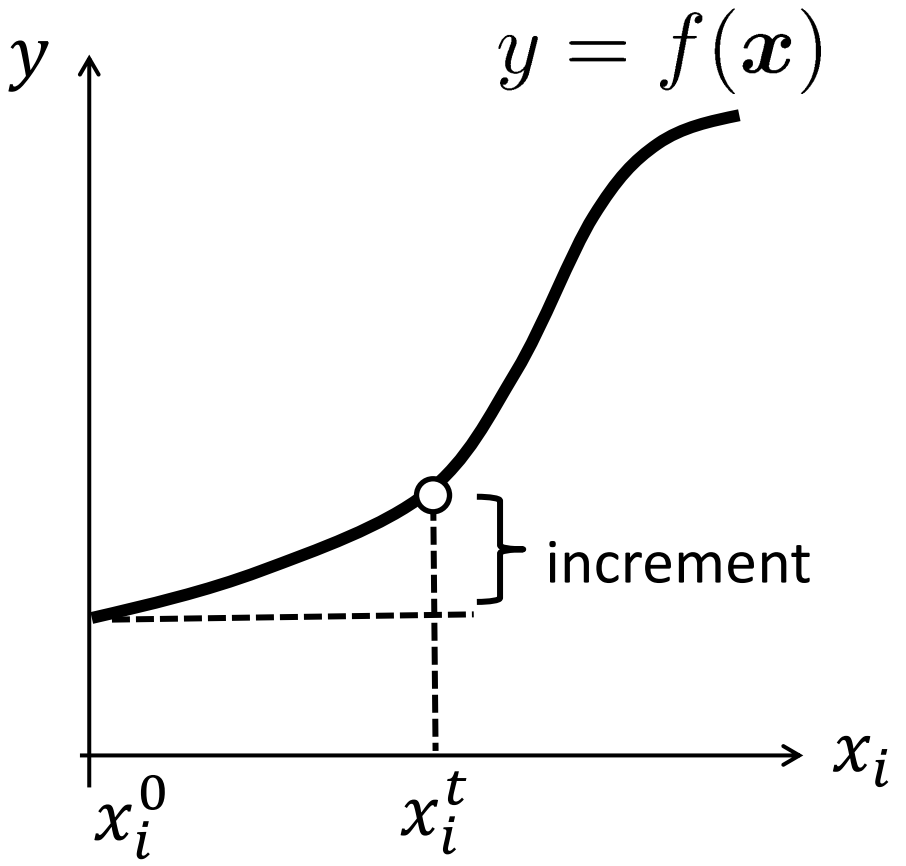
- John Sipple. "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure," In Proceedings of the 37th International Conference on Machine Learning (ICML 20).

Expected IG (EIG) eliminates IG's 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] and uses a certain Monte Carlo sampling.

- 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:

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

$f(\cdot)$'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

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

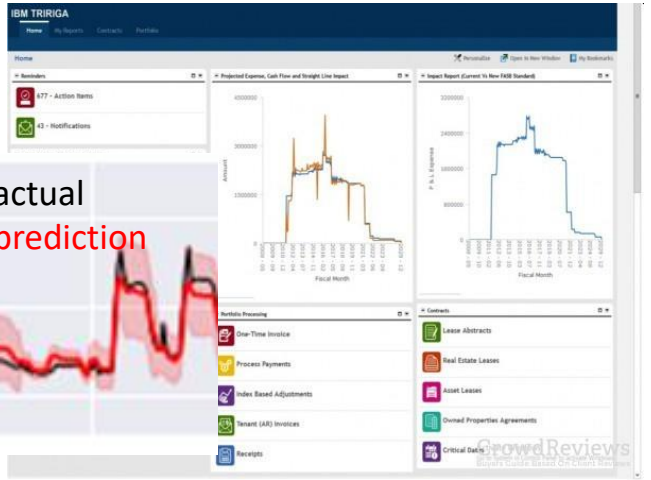
Idé, T., & Abe, N. (2023, August). Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 845-856).

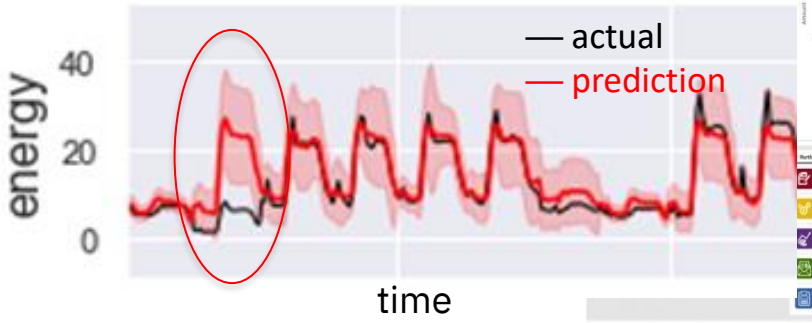
Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
- Data science problems in semiconductor manufacturing

Motivating problem: “Is there any issue with my building’s A/C system?”

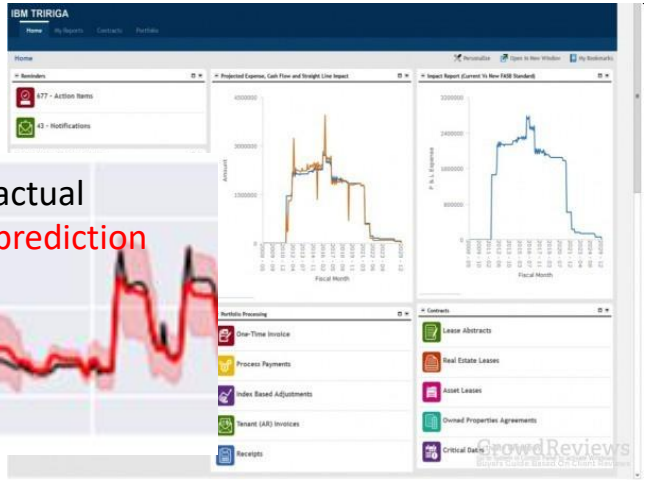
- IBM TRIRIGA is a software platform of smart building management.
  - Energy consumption monitoring is a key feature.
- Situation:
  - A building owner noticed a significant difference between forecasted and actual energy consumption.
  - Suspecting a potential issue with the A/C system, the owner consulted IBM for advice.
- Question:
  - What kind of data science problem does this situation suggest?

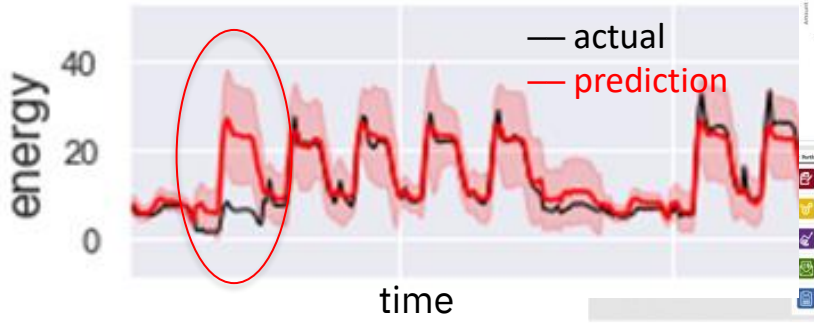




Motivating problem: “Is there any issue with my building’s A/C system?”

- What kind of data science problem does this situation suggest?
- Typical constraints:
  - You have access to input and output data, but not the details of the forecasting algorithm.
  - You cannot definitively pinpoint the root cause due to many unknown factors.
- Solving the attribution problem would probably be the safest option for you as a data scientist.





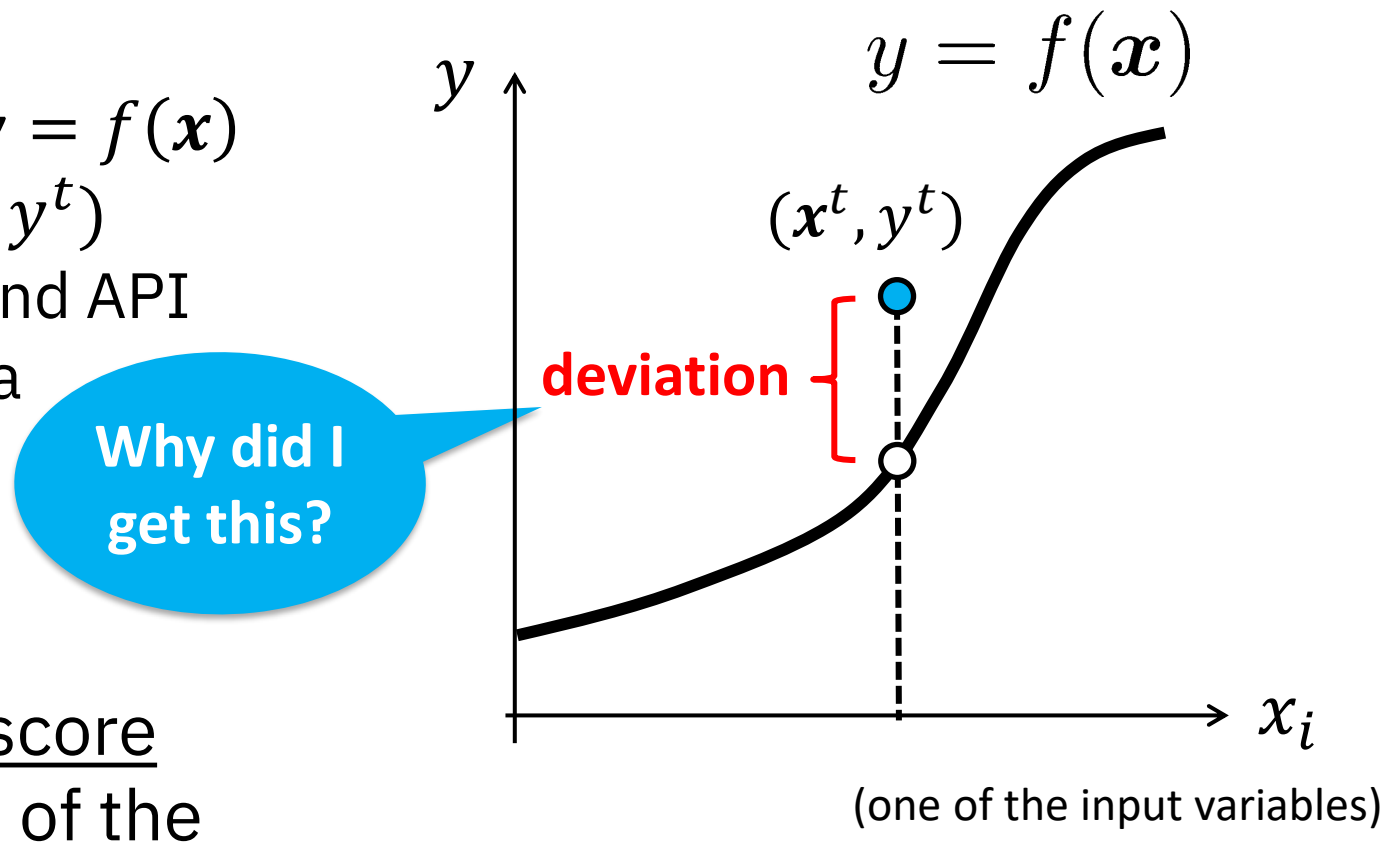
The anomaly attribution problem: Explaining deviations between prediction and observation

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

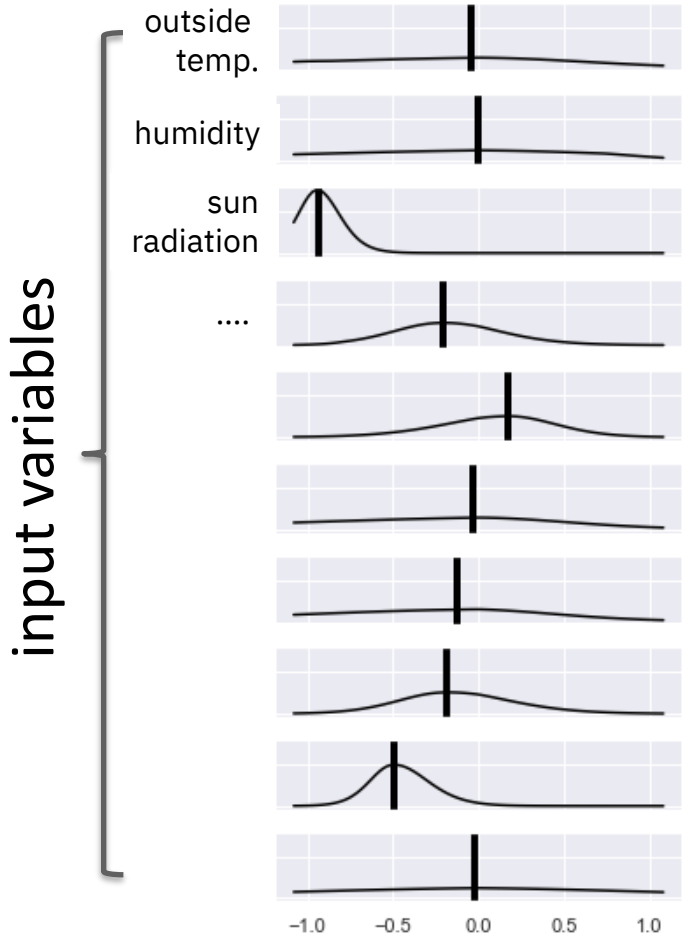
Explain:

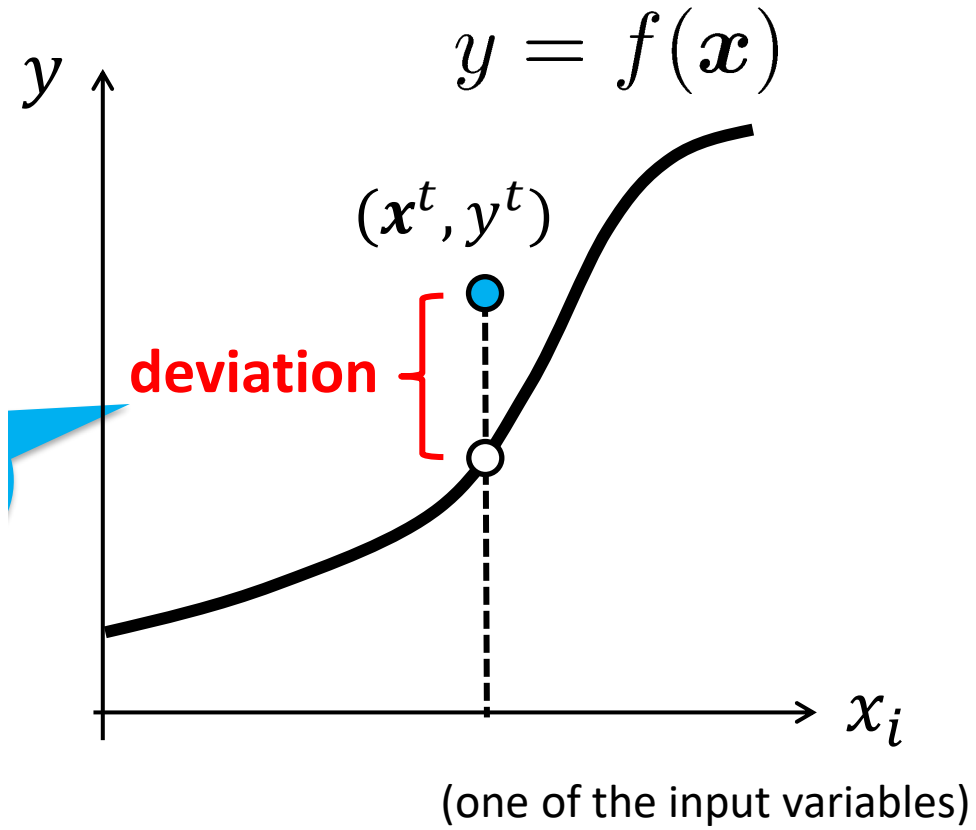
- The deviation $f(\boldsymbol{x}^t) - y^t$
- by computing the attribution score (responsibility score) for *each* of the input variables $\boldsymbol{x}$.



The anomaly attribution problem: Explaining deviations between prediction and observation

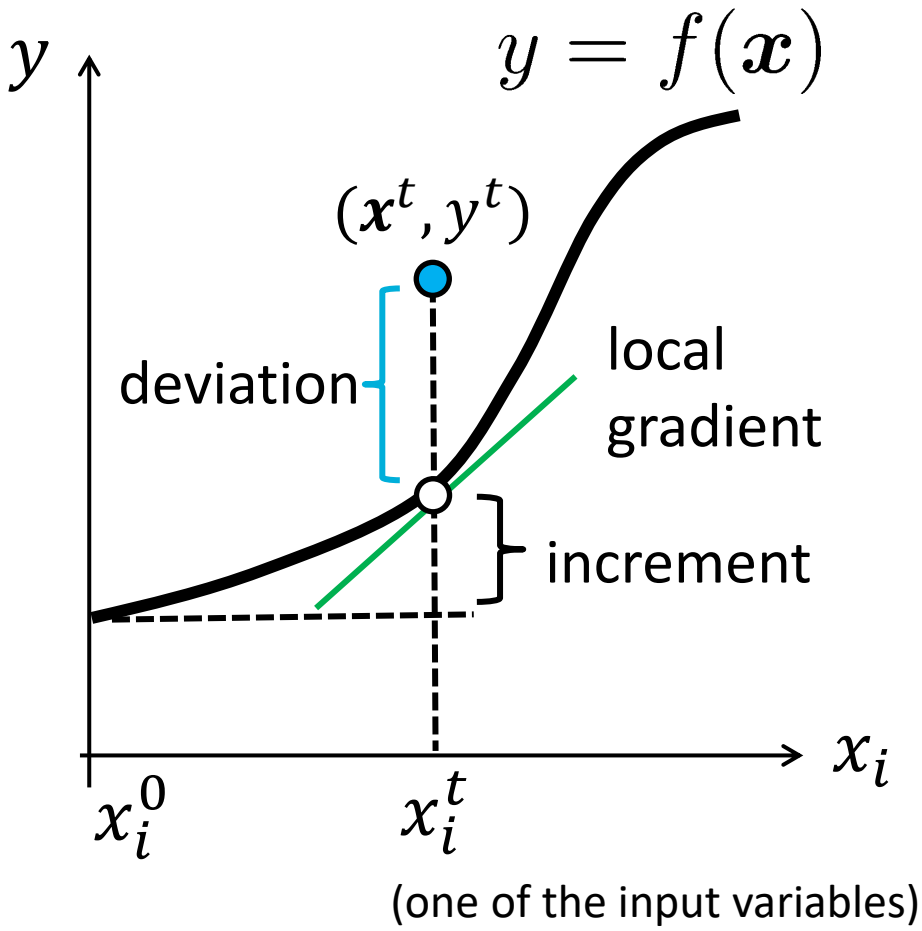
Hopefully, we want to get the score's confidence as well.





Anomaly attribution is not the same as the standard attribution task.

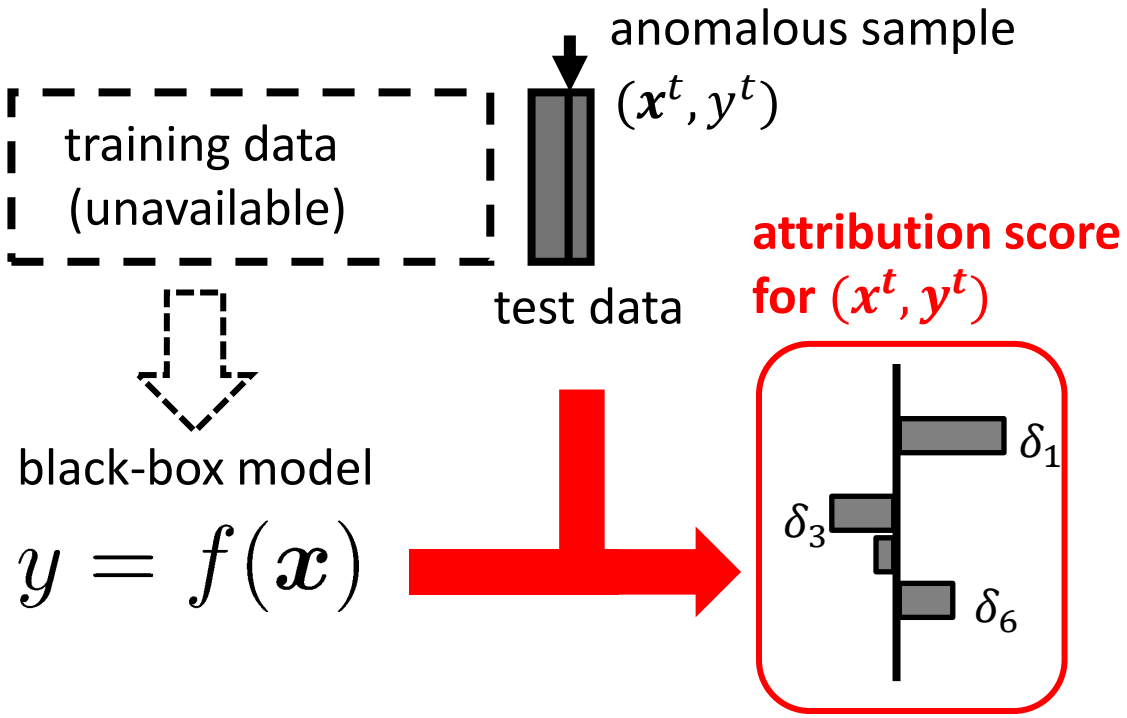
- LIME, SV, IG, and EIG are deviation-agnostic.
  - These methods explain $f(\boldsymbol{x})$ locally at $\boldsymbol{x} = \boldsymbol{x}^t$, independently of the observed outcome value y.
  - The limitation remains true even if we aim to explain the function $f(\boldsymbol{x}) - y$ rather than just $f(\boldsymbol{x})$.
- 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.

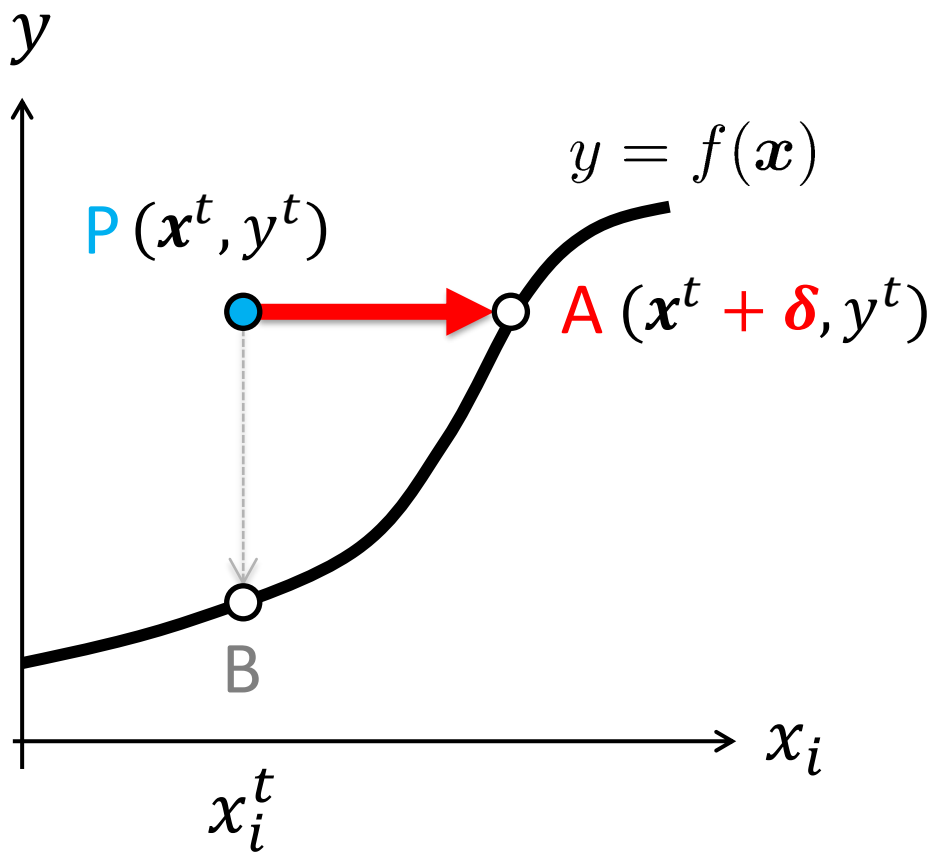
Seeking a new approach beyond local linear approximation

- Developed a new attribution algorithm so it can provide deviation-sensitive attribution scores.
- The new approach is based on the concept of counterfactual perturbation.
  - → next page



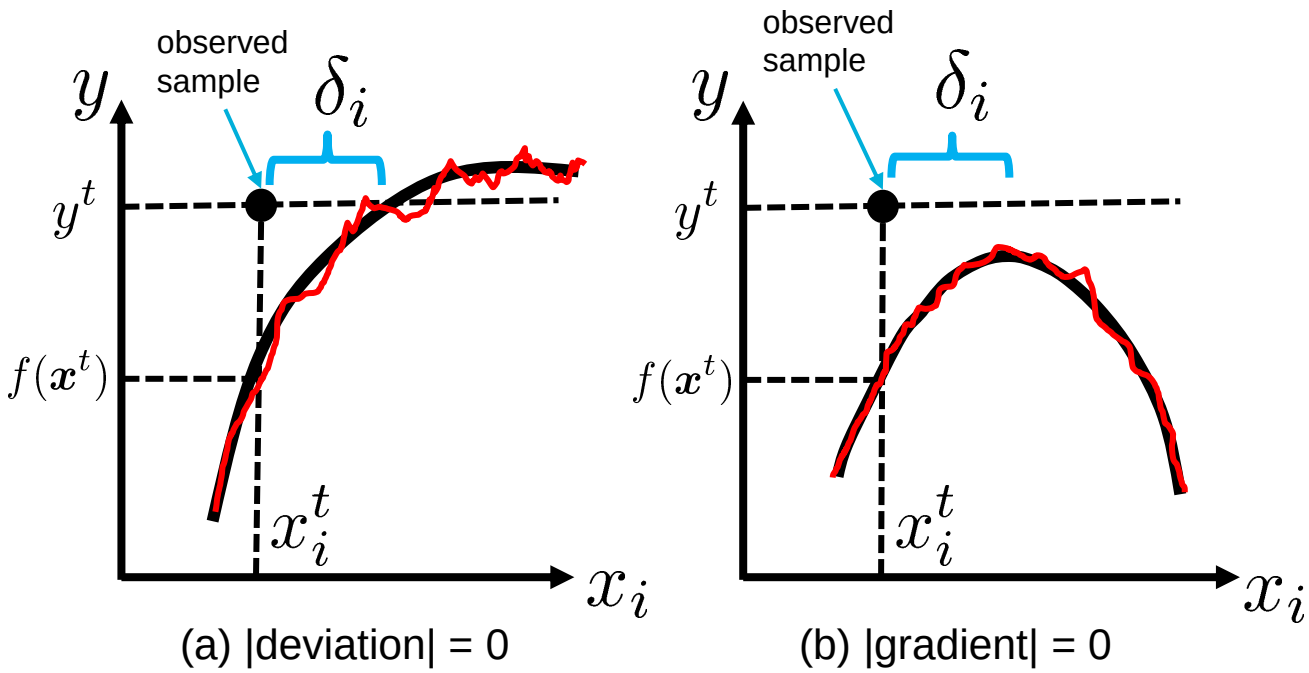
Given a test point (x^t, y^t) being anomalous, we ask: How much "work" would we need to bring it to the normalcy?

- 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.
- Use $\boldsymbol{\delta}$ as the attribution score.



Perturbation as explanation: Technical notes of the Likelihood Compensation (LC) algorithm

- LC uses the likelihood to find an optimal perturbation.
  - This makes LC less ad hoc than methods like LIME, as it leverages the likelihood used during model training.
- LC needs to solve an optimization problem to determine $\boldsymbol{\delta}$.
  - LC seeks the point with either zero deviation or zero gradient.
    ✓ →
- LC can be extended to evaluate the uncertainty of the score.

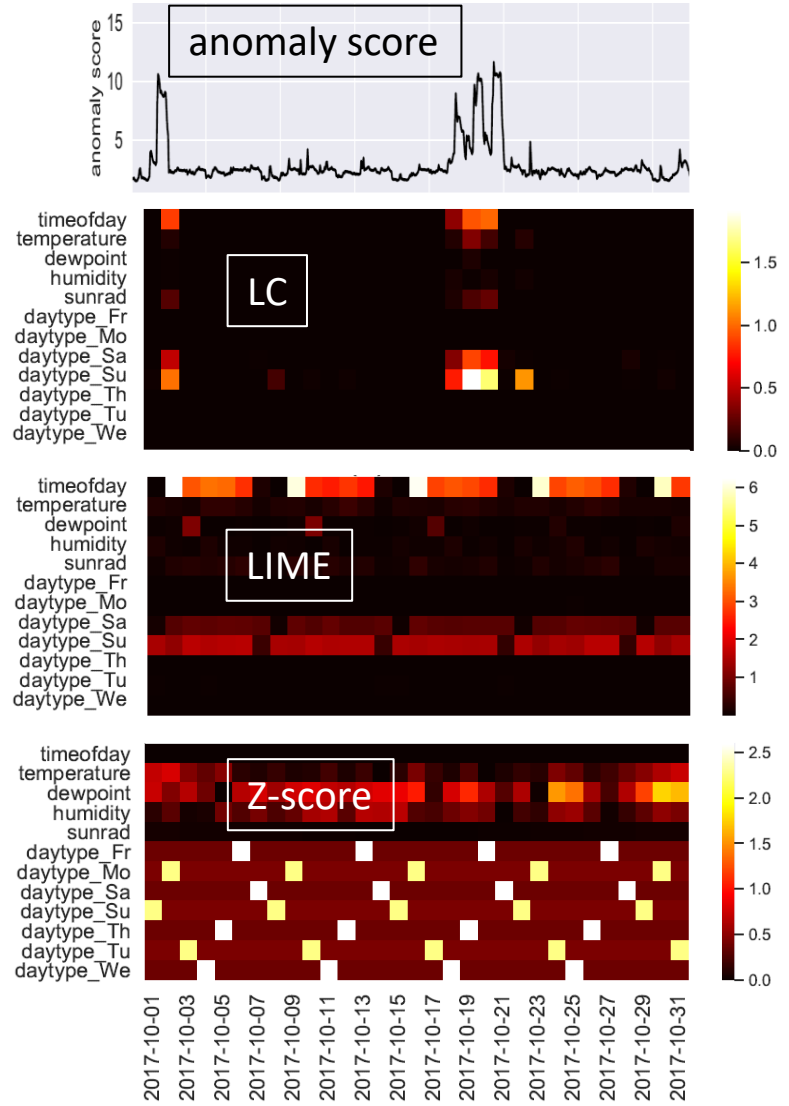


- T. Idé, et al., "Anomaly Attribution with Likelihood Compensation," In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 21, February 2-9, 2021, virtual), pp.4131-4138
- T. Idé, N. Abe, "Generative Perturbation Analysis for Probabilistic Black-Box Anomaly Attribution,"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

Motivating problem: “Is there any issue with my building’s A/C system?”

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



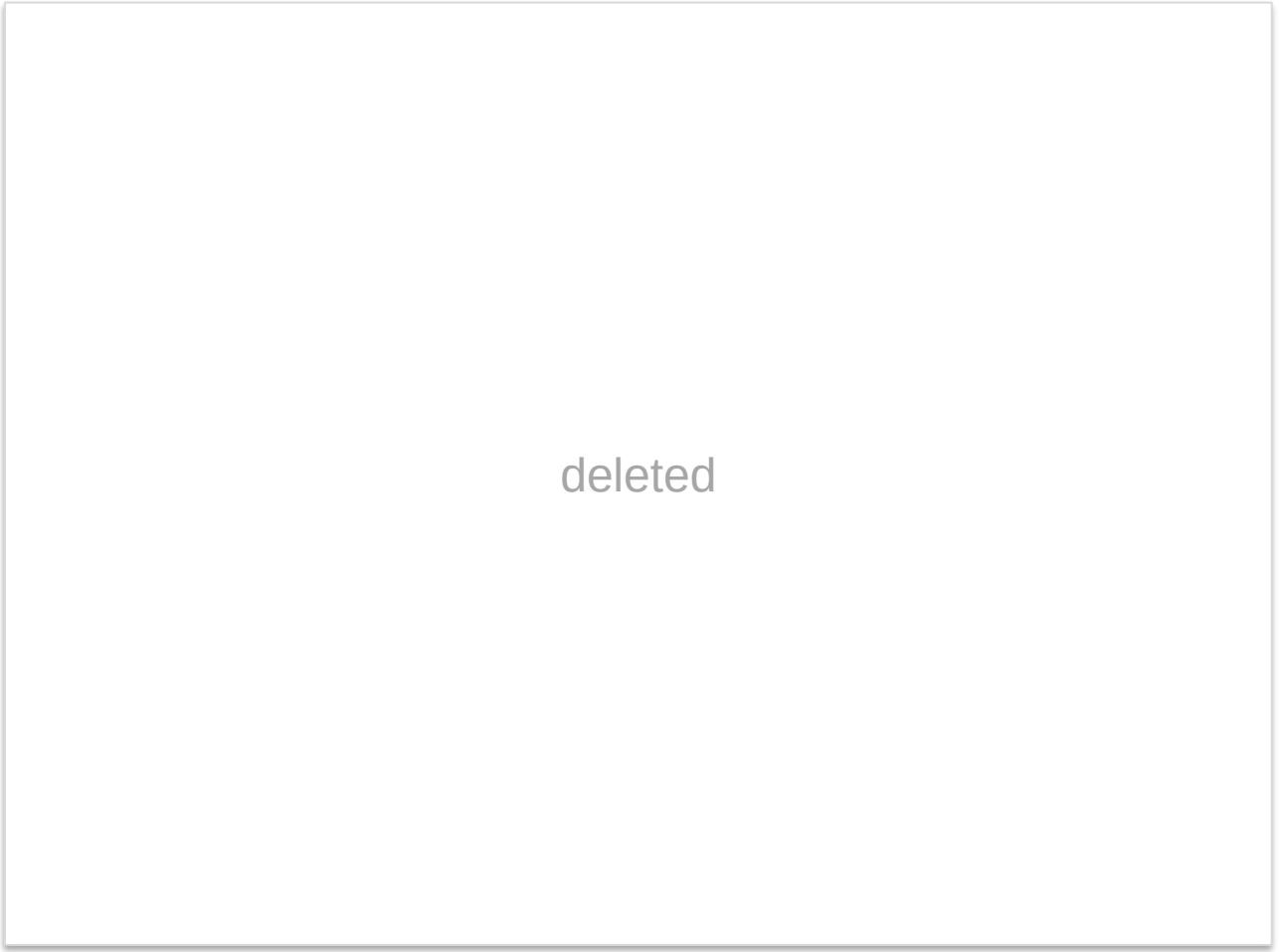
Agenda

- Introduction to explainable AI (XAI)
- Existing local attribution methods for regression
- The anomaly attribution problem
- Data science problems in semiconductor manufacturing

Semiconductor manufacturing: A rich field for data science applications

- Wafers are processed according to a predefined route, yet significant variations arise even with identical routes.

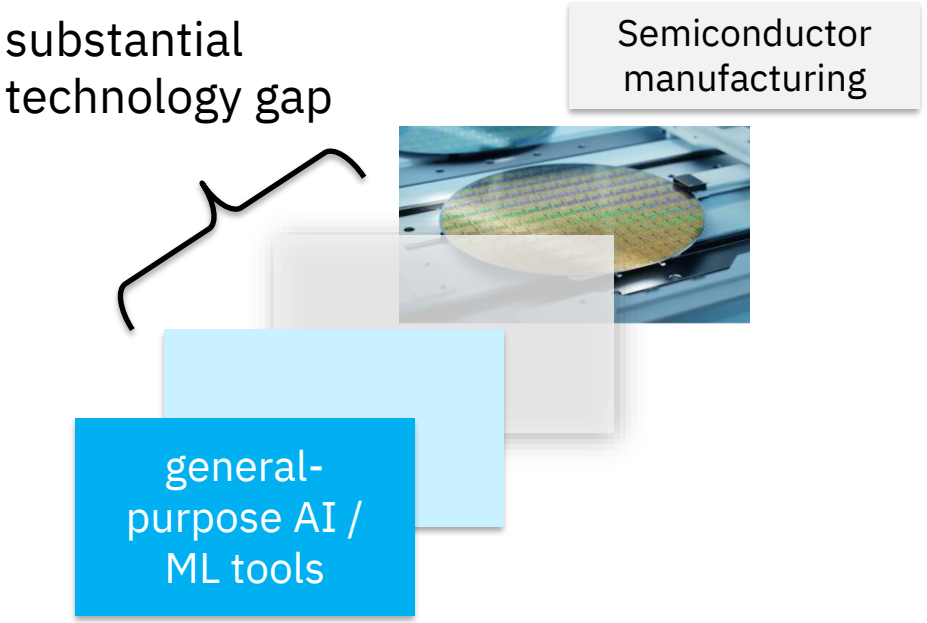
- Sources of variation
  - Duplicated tools and chambers with subtle characteristic differences.
  - Variable waiting times in time-sensitive processes.
  - Tool degradations over time.
  - Ad hoc adjustments to processing recipes.
  - etc.
- Statistical machine learning is crucial to handle the process variabilities to enhance production yield.



Source: Steve Blank, "The Semiconductor Ecosystem – Explained," https://steveblank.com/2022/01/25/the-semiconductor-ecosystem/

Fact: General-purpose ML tools rarely drive significant semiconductor innovation

- Despite overwhelming total data volume, the **effective sample size** can be small.
  - One possible definition:
    - ✓ $\dfrac{(\#\ \text{of wafers})}{(\#\ \text{of distinct processing conditions})}$
- Each wafer undergoes 1000s of processing steps with a slightly different condition in each step
  - Slight variations in conditions at each step lead to an enormous number of similar yet distinct processing routes.

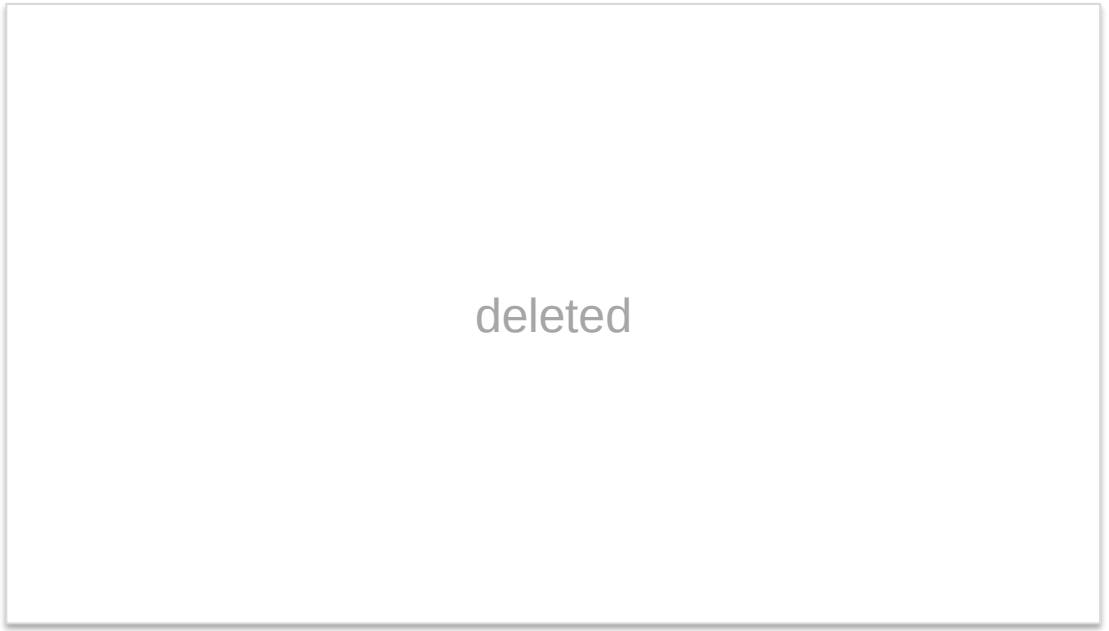


Did you know …?

Average wafer count of distinct processing routes in an R&D fab: ~1

Predicting a human life ≈ Predicting wafer outcomes?
One scene from Netflix film “Don’t Look Up” (2021)

- Dr. Mindy (Leonardo DiCaprio) is challenged by an excentric billionaire, Peter Isherwell.
- Do you think Peter’s claim is plausible?
- Wafers
  - Process-process dependency.
  - Process variations affecting in a combinatorial fashion.
- Humans
  - A major or minor event can profoundly alter a person’s trajectory.
  - Random and uncontrollable elements (e.g., chance encounters, genetics, etc.) can decisively influence a person's future.



Peter: “You know that BASH has over 40 million data points on you … I know what you are.”

“Our algorithms can even predict how you'll die. To 96… 96.5% accuracy.”

(quote source: IMDB)

Data science problems in semiconductor manufacturing

deleted

deleted

Thank you!

Explaining what went wrong: The black-box anomaly attribution problem

- Abstract:
  - Explainable AI (XAI) is one of the hottest topics in machine learning research. In real business and industrial applications, one particularly important scenario is explaining the gap between a prediction and the actual outcome, as this gap may indicate some system issue if the model has been trained on normal data. Attribution is the task that provides explanations by quantifying the responsibility of each input variable.
  - In this talk, I will first review popular attribution approaches, including linear surrogate modeling (LIME), Shapley values, and integrated gradients. Next, I will discuss the limitations of these algorithms and how I addressed these challenges in a real-world anomaly attribution use case involving building energy management. If time permits, I will also share real-world anomaly attribution tasks from semiconductor manufacturing at IBM.
- Bio:
  - Dr. Tsuyoshi ("Ide-san") Ide is 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.