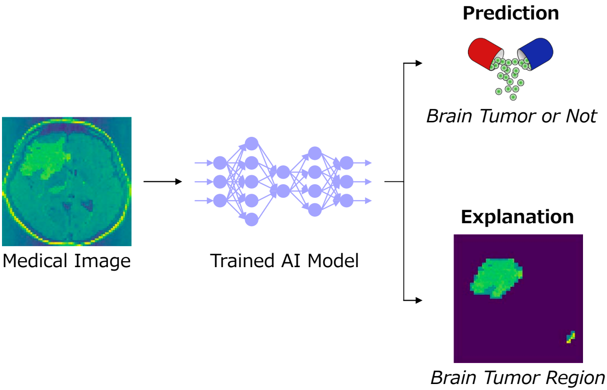
Statistical Test for Explainable AI

Ichiro Takeuchi (Data-driven Bio-medical Science Team)

This is joint work with D. Miwa and V.N.L. Duy, and will be presented at ICLR2023

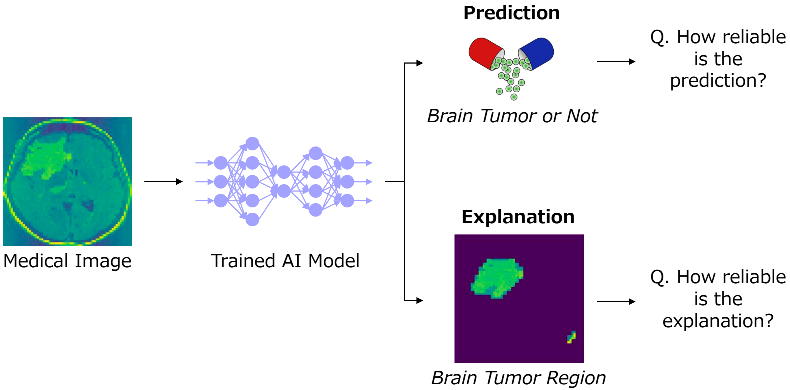
AI for Science: Prediction and Explanation

- Two goals in "AI for Science": *prediction* and *explanation*.



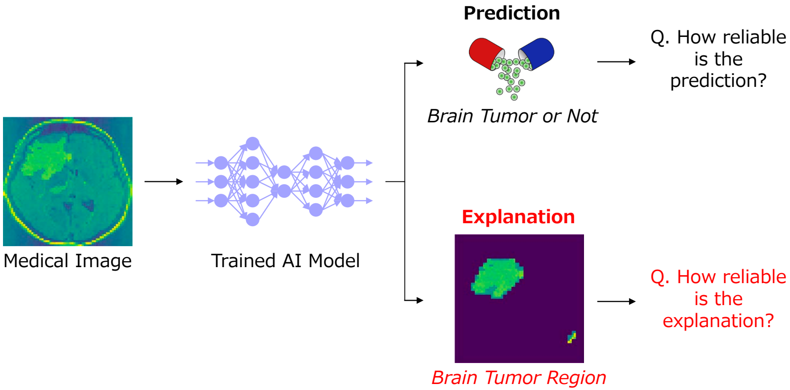
Reliability in AI-Driven Science

- Quantifying the reliability of AI-driven predictions and discoveries is required.



Reliability in AI-Driven Science

- Quantifying the reliability of AI-driven predictions and discoveries is required.



How can we quantify the reliability of explanation?

- Consider a linear model case:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$$

- Suppose we have the following parameter estimation result:

$$\hat{\beta}_3 = 3.4$$

- Statistical test for the coefficient β_3

$$\mathrm{H}_0 : \beta_3 = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \beta_3 \neq 0$$

- Statistical significance measures: p-values

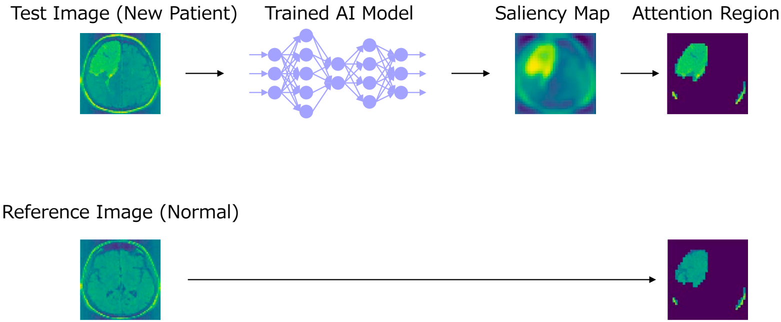
$$p_3 = \mathrm{Pr}_{\mathrm{H}_0}\left(|\hat{\beta}_3| \geq 3.4\right)$$

- Interpretation (with the significance level, e.g., $\alpha = 0.05$)

$$\begin{aligned} p_3 < 0.05 \quad &\Rightarrow \quad x_3 \text{ is a reliable explainable feature} \\ p_3 \geq 0.05 \quad &\Rightarrow \quad x_3 \text{ is not a reliable explainable feature} \end{aligned}$$

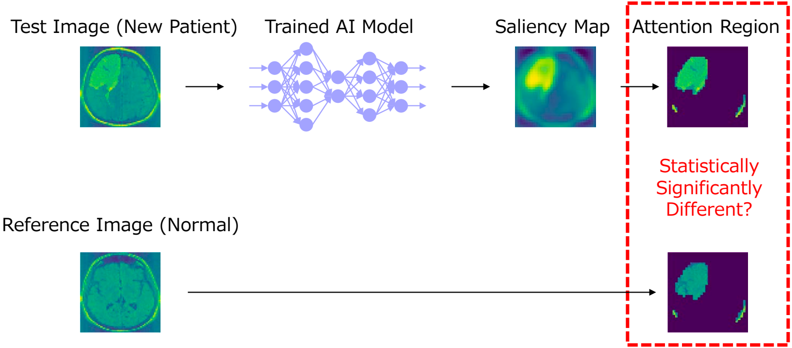
Statistical Testing Framework for AI-Driven Hypotheses

- Consider quantifying reliability in the framework of statistical tests.



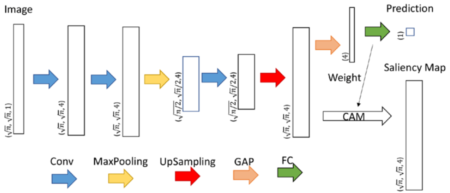
Statistical Testing Framework for AI-Driven Hypotheses

- Consider quantifying reliability in the framework of statistical tests.



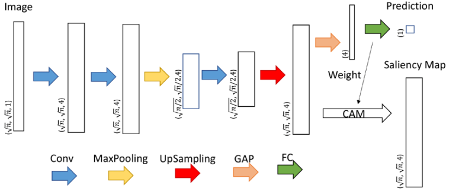
A Working Example

Step 1. We trained a neural network with training set, which includes 939 images with tumors and 941 images without tumor:

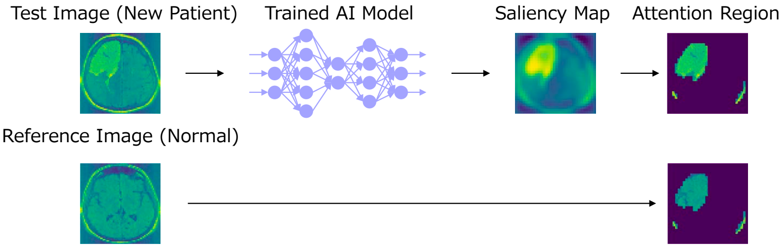


A Working Example

Step 1. We trained a neural network with training set, which includes 939 images with tumors and 941 images without tumor:

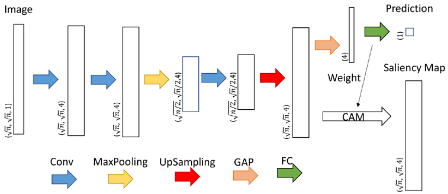


Step 2. We input several test images to the trained network and conduct naïve two-sample test without caring that the attention region is obtained by the network.

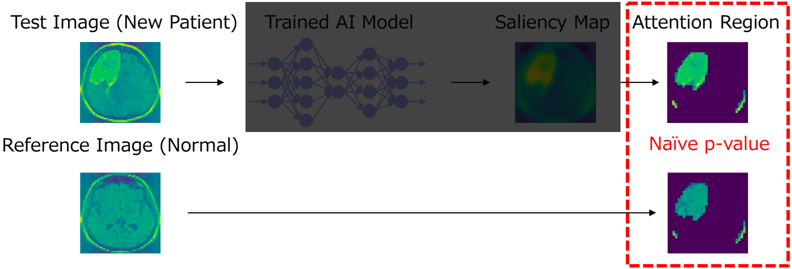


A Working Example

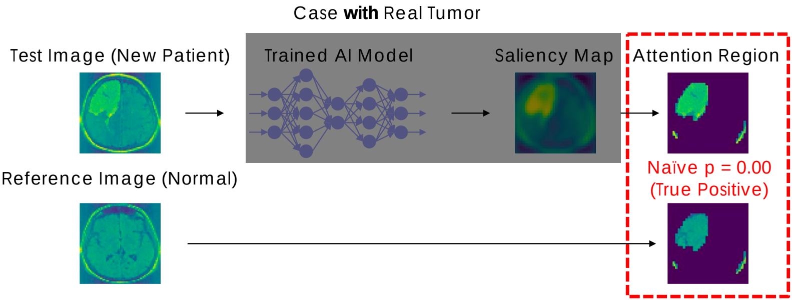
Step 1. We trained a neural network with training set, which includes 939 images with tumors and 941 images without tumor:



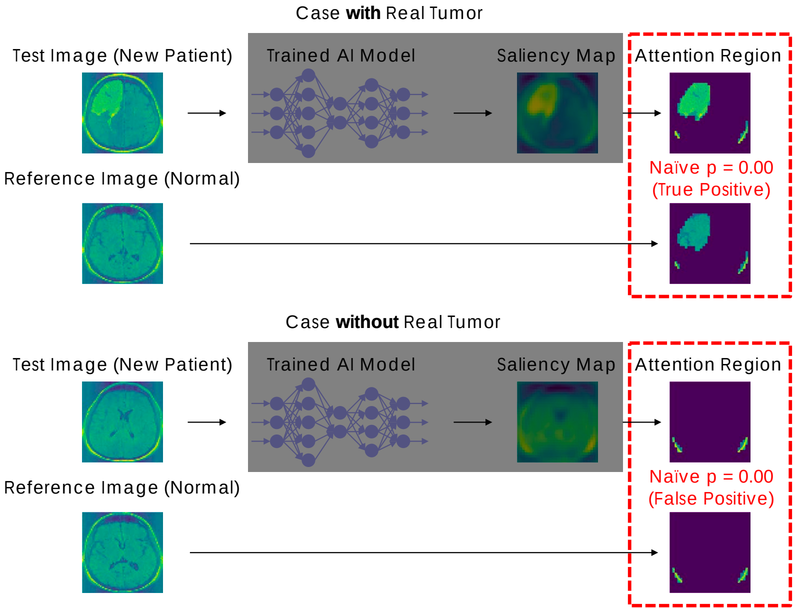
Step 2. We input several test images to the trained network and conduct naïve two-sample test without caring that the attention region is obtained by the network.



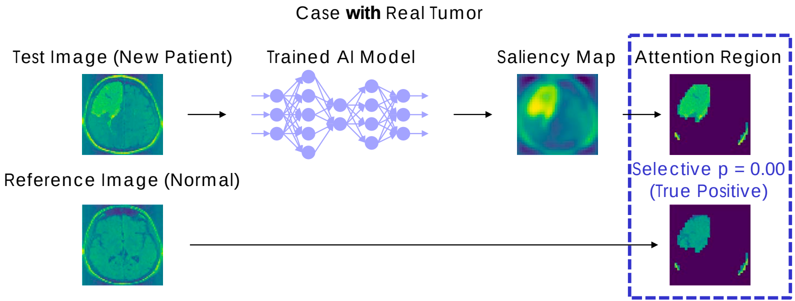
Traditional Statistical Inference: Naive p-Values



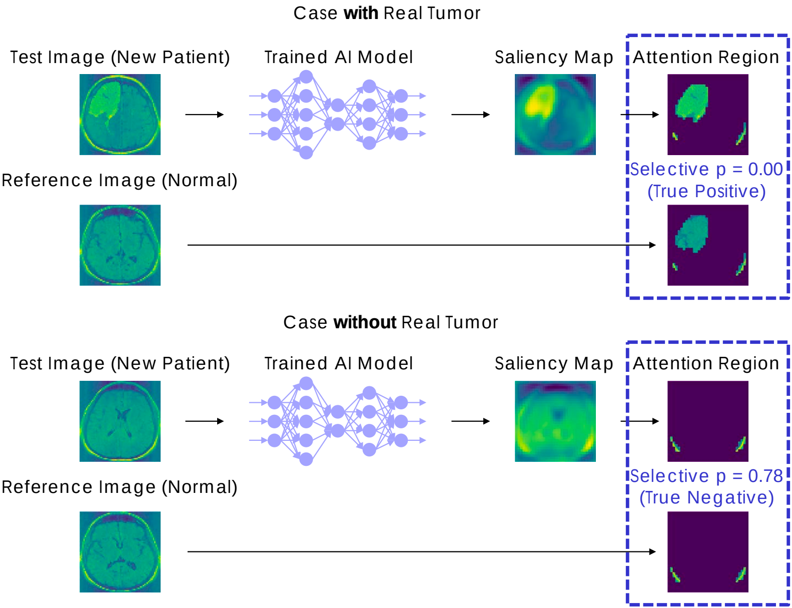
Traditional Statistical Inference: Naive p-Values



Selective Inference (Statistical Inference for Data-Driven Hypotheses): S



Selective Inference (Statistical Inference for Data-Driven Hypotheses): S



The Messages in This Talk

- Why naive p-values are invalid for AI-driven hypotheses and how we interpret and formulate this issue?

- How selective inference (a new trend in statistics for data-driven hypotheses) resolve this issue?

- How we can compute selective p-values for deep neural network-driven hypotheses?

The Messages in This Talk

- Why naive p-values are invalid for AI-driven hypotheses and how we interpret and formulate this issue?
- How selective inference (a new trend in statistics for data-driven hypotheses) resolve this issue?
- How we can compute selective p-values for deep neural network-driven hypotheses?

The Messages in This Talk

- Why naive p-values are invalid for AI-driven hypotheses and how we interpret and formulate this issue?

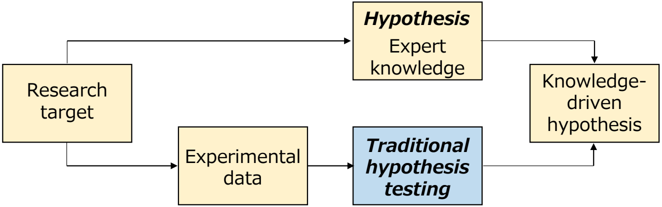
- How selective inference (a new trend in statistics for data-driven hypotheses) resolve this issue?

- How we can compute selective p-values for deep neural network-driven hypotheses?

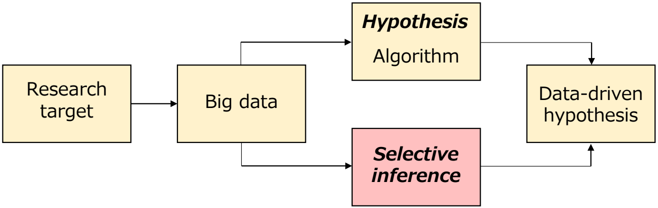
Hypothesis Selection Bias and Multiple Comparison

Knowledge-Driven vs. Data-Driven Science

(Traditional) Knowledge-driven science

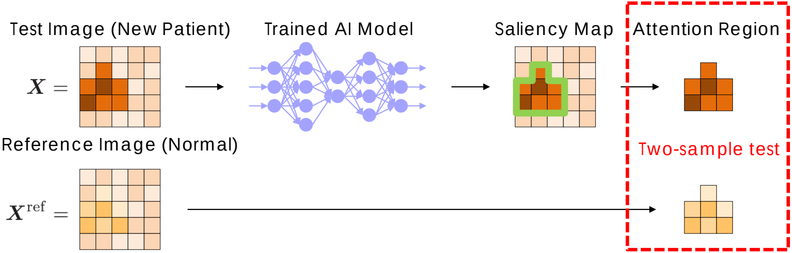


Data-driven science



Problem Formulation

- Goal: Identify the attention region in a medical image by a saliency method (e.g., CAM).



- An image is represented as an n-dimensional vector of pixel values $X \in \mathbb{R}^n$ as

$$\text{Test Image:} \quad \underbrace{X}_{\text{image}} = \underbrace{s}_{\text{signal}} + \underbrace{\varepsilon}_{\text{noise}}, \qquad \underbrace{\varepsilon \sim \mathbb{N}(\mathbf{0}, \Sigma)}_{\text{Normally-distributed noise}}$$

$$\text{Reference Image:} \quad \underbrace{X^{\text{ref}}}_{\text{image}} = \underbrace{s^{\text{ref}}}_{\text{signal}} + \underbrace{\varepsilon^{\text{ref}}}_{\text{noise}}, \qquad \underbrace{\varepsilon^{\text{ref}} \sim \mathbb{N}(\mathbf{0}, \Sigma)}_{\text{Normally-distributed noise}}$$

- Algorithm (Trained Network) $\mathcal{A}$

$$\underbrace{\mathcal{A}}_{\text{algorithm}} : \underbrace{X}_{\text{image}} \mapsto \underbrace{\mathcal{M}_X}_{\text{attention region}}$$

Hypothesis Testing

- Mean Null Test
  - Null Hypothesis H_0 and Alternative Hypothesis H_1

$$\mathrm{H}_0 : \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} s_i = \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} s_i^{\mathrm{ref}} \quad \text{vs.} \quad \mathrm{H}_1 : \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} s_i \neq \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} s_i^{\mathrm{ref}}$$

  - Test statistic

$$\Delta_X := \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} X_i - \frac{1}{|\mathcal{M}_X|}\sum_{i\in\mathcal{M}_X} X_i^{\mathrm{ref}}$$

- Global Null Test
  - Null Hypothesis H_0 and Alternative Hypothesis H_1

$$\mathrm{H}_0 : s_i = s_i^{\mathrm{ref}}\ \forall i \in \mathcal{M}_X \quad \text{vs.} \quad \mathrm{H}_1 : s_i \neq s_i^{\mathrm{ref}}\ \exists i \in \mathcal{M}_X$$

  - Test-statistic

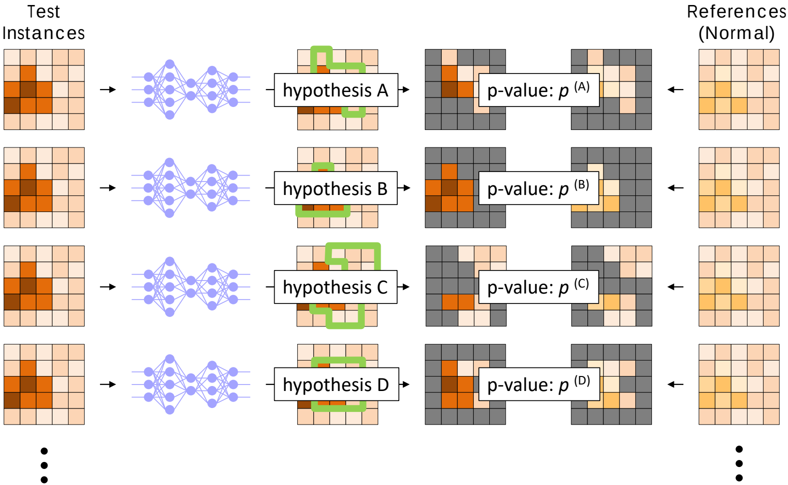
$$\Delta_X = \sqrt{\sum_{i\in\mathcal{M}_X}\left(\frac{X_i - X_i^{\mathrm{ref}}}{\sqrt{2}\sigma}\right)^2}$$

- Statistical significance (two-sided p-value)

$$p = \Pr\left(\underbrace{|\Delta_X|}_{\text{random variable}} \geq \underbrace{|\Delta_{X_{\mathrm{obs}}}|}_{\text{observation}}\right)$$

Multiple Testing / Hypothesis Selection Interpretation

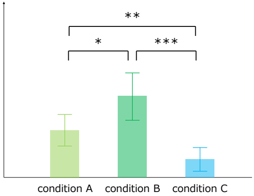
- The data-driven hypothesis is interpreted as the result of multiple comparison with all possible $2^{\#\text{pixels}}$ results.



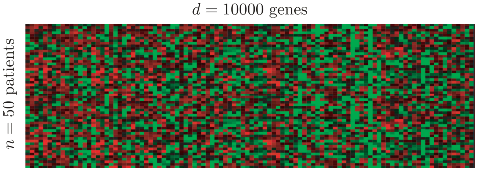
- Correction of the selection bias is indispensable in multiple comparison.

Multiple Comparison

- In the context of traditional multiple hypothesis testing, only a handful of tests are considered.



- In the context of genetic data analysis (2000∼), large-scale multiple comparison with tens of thousands of hypotheses were considered.



- The number of all possible hypotheses that AI/ML can produce is much more than the existing methods can handle.

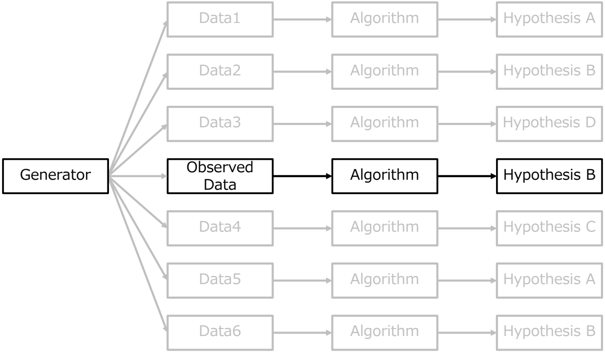
Three approaches for multiple comparison correction

- **Family-wise error rate (FWER) control**: controlling the probability of finding a false positive (FP) $< \alpha$ (e.g., 0.05)
- **False discover rate (FDR)**: controlling the expected proportion of discoveries that are false $< \alpha$ (e.g., 0.05)
- **Conditional selective inference (SI)**: controlling the probability of finding a FP conditional on the hypothesis selection event $< \alpha$ (e.g., 0.05)

Conditional Selective Inference (SI)

Basic idea of conditional SI

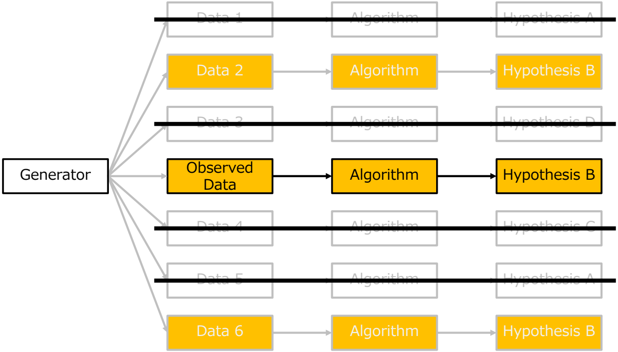
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness where the same hypothesis is selected, the hypothesis selection bias disappears.

Basic idea of conditional SI

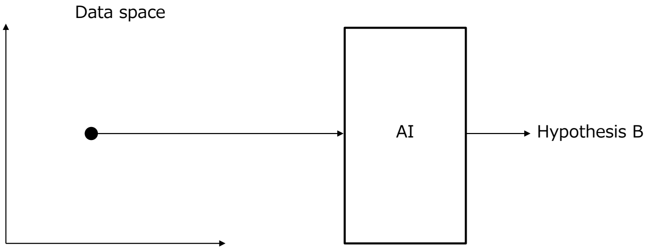
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness where the same hypothesis is selected, the hypothesis selection bias disappears.

Basic idea of conditional SI

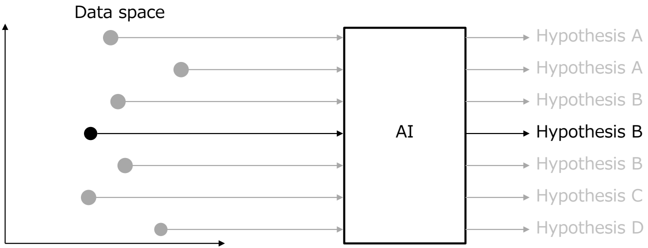
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

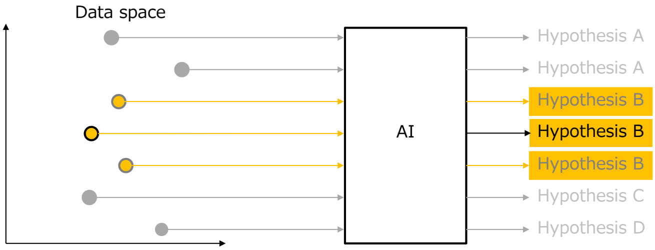
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

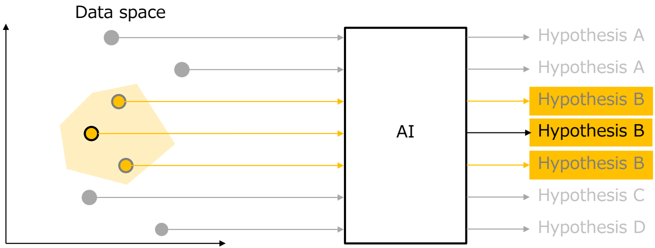
- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Basic idea of conditional SI

- The key idea of conditional SI is to consider only the cases (parallel worlds) where the same hypothesis is selected.



- Intuitively, by considering only the randomness in the subset of the data space, the hypothesis selection bias disappears.

Conditional SI for The Working Problem

- Ordinary statistical significance (p-value)

$$p = \Pr\left(\underbrace{|\Delta_X|}_{\text{random var.}} \geq \underbrace{|\Delta_{X_{\rm obs}}|}_{\text{observation}} \right)$$

- Conditional statistical significance (selective p-value)

$$p = \Pr\left(\underbrace{|\Delta_X|}_{\text{random var.}} \geq \underbrace{|\Delta_{X_{\rm obs}}|}_{\text{observation}} \;\middle|\; \underbrace{\mathcal{M}_X = \mathcal{M}_{X_{\rm obs}}}_{\text{the same attention region is selected}} \right)$$

Ordinary p-values vs. Selective p-values

- The ordinary p-values are too complicated to compute for data-driven hypotheses obtained by complicated algorithms.

- The selective p-values are computable as long as the selection event of the selected hypotheses are characterized in tractable way.

- The key idea of conditional SI is to decouple the "hypothesis selection" and "statistical inference" so that the latter can be done as if the hypothesis is fixed.

History of Conditional SI Research

- The notion of conditional inference has long been used in many problems and known in the literature of statistics.
- Lee et al. [1] first proposed a computationally tractable conditional SI method (Polyhedron-based SI) for Lasso.
- Inspired by this work, polyhedron-based SI for various other feature selection methods were developed (e.g., [2, 3, 4, 5]).
- Polyhedron-based SI has been found useful for statistical inference of various data-driven hypotheses other than feature selection (e.g., [6, 7, 8, 9])
- Conditional SI loses its power by *over-conditioning* — several approaches have begun to be studied to resolve this problem (e.g., [10, 11, 12, 13]).

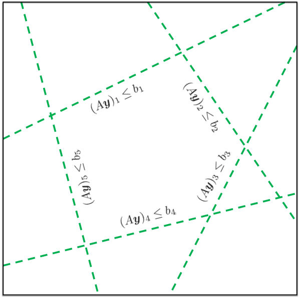
Selective Inference for Lasso [1]

- Lee et al. [1] developed a SI framework when the selection event is characterized by a set of linear inequalities in the form of

$$Ay \leq b \quad \text{(for a certain matrix } A \text{ and a vector } b\text{)},$$

and found that the selection event for Lasso ($\mathcal{A}_{\text{Lasso}}$) can be fit into this framework:

$$\{\text{“selected features”} \leftarrow \mathcal{A}_{\text{Lasso}}(y)\} \quad \Leftrightarrow \quad Ay \leq b.$$



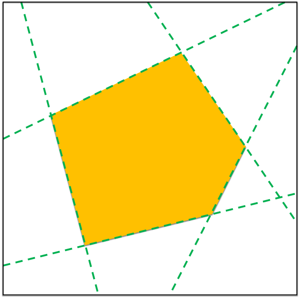
Selective Inference for Lasso [1]

- Lee et al. [1] developed a SI framework when the selection event is characterized by a set of linear inequalities in the form of

$$Ay \leq b \quad \text{(for a certain matrix } A \text{ and a vector } b\text{)},$$

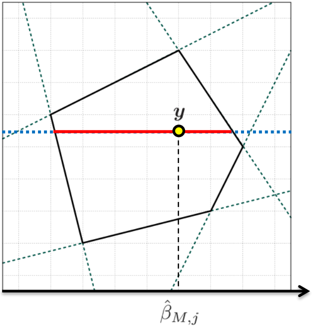
and found that the selection event for Lasso ($\mathcal{A}_{\mathrm{Lasso}}$) can be fit into this framework:

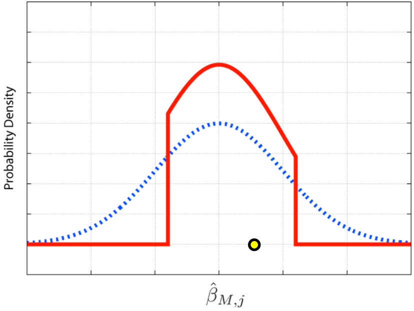
$$\{\text{“selected features”} \leftarrow \mathcal{A}_{\mathrm{Lasso}}(y)\} \quad \Leftrightarrow \quad Ay \leq b.$$



Truncated Normal Distribution for Polyhedron-based SI

- When $y \sim N(\mu, \Sigma)$ and the selection event is characterized by a polyhedron, the conditional sampling distribution of $\hat{\beta}_{\mathcal{M},j}$ is in the form of truncated Normal distribution.

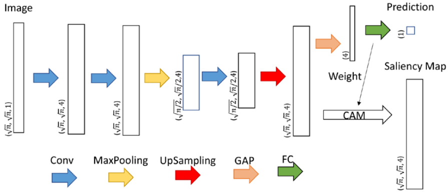




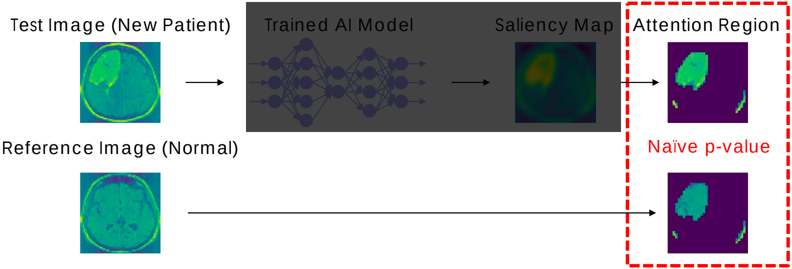
Conditional Selective Inference for Deep Learning

Problem Setup (Revisited)

Step 1. We trained a neural network with training set, which includes 939 images with tumors and 941 images without tumor:

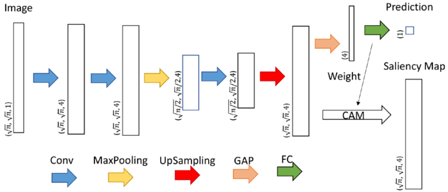


Step 2a. We input several test images to the trained network and conduct naïve two-sample test without caring that the attention region is obtained by the data.

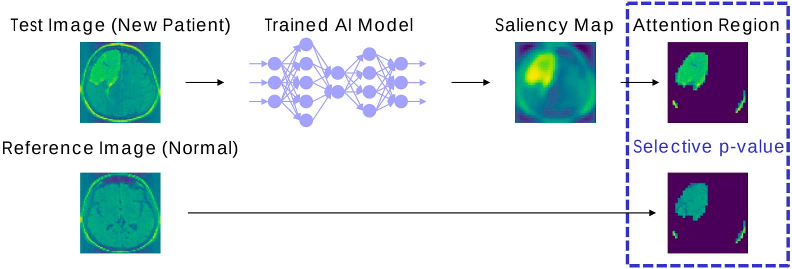


Problem Setup (Revisited)

Step 1. We trained a neural network with training set, which includes 939 images with tumors and 941 images without tumor:

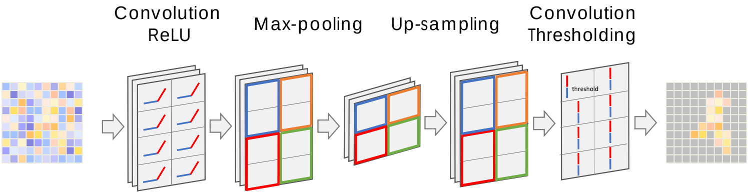


Step 2b. We input several test images to the trained network and conduct selective two-sample test by properly caring that the attention region is obtained by the data.



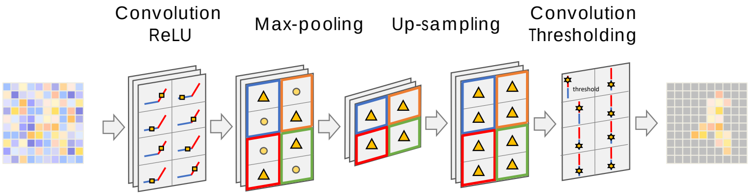
Piecewise-Linear Network

- Most of the components in convolutional neural network (CNN) can be represented or precisely approximated as piecewise-linear functions.



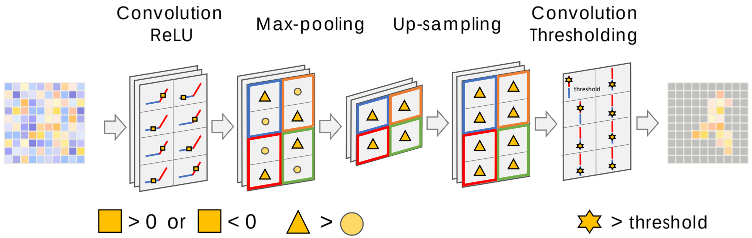
Piecewise-Linear Network

- Most of the components in convolutional neural network (CNN) can be represented or precisely approximated as piecewise-linear functions.



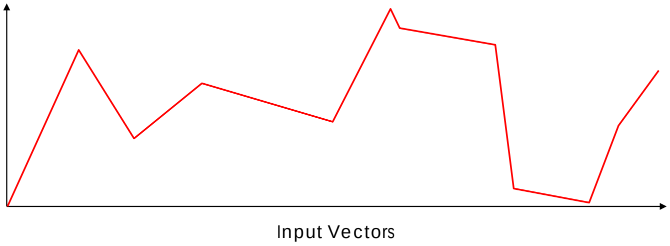
Piecewise-Linear Network

- Most of the components in convolutional neural network (CNN) can be represented or precisely approximated as piecewise-linear functions.



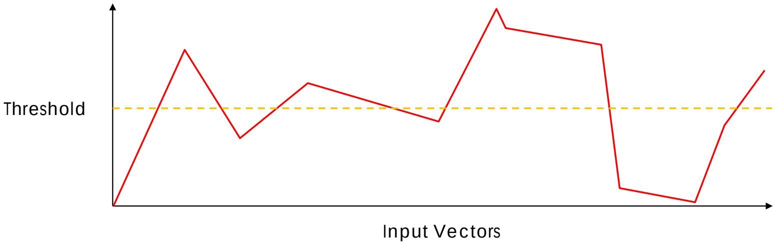
Selection Event by Piecewise-Linear Functions

- A selection event characterized by finite number of piecewise-linear functions looks like:



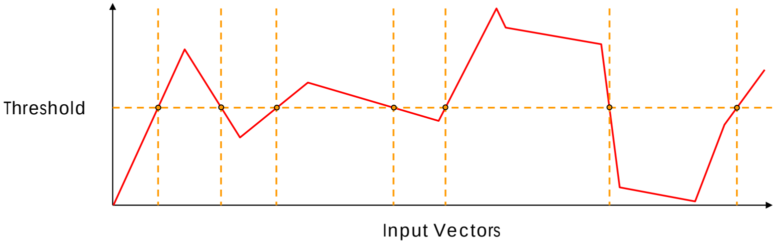
Selection Event by Piecewise-Linear Functions

- A selection event characterized by finite number of piecewise-linear functions looks like:



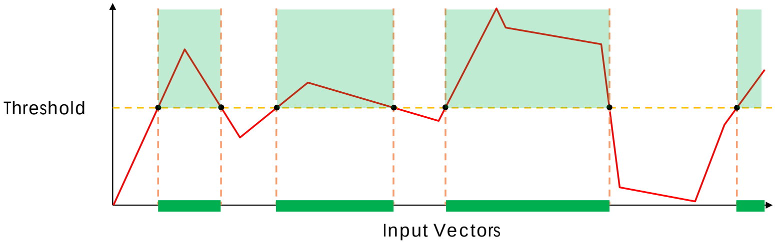
Selection Event by Piecewise-Linear Functions

- A selection event characterized by finite number of piecewise-linear functions looks like:



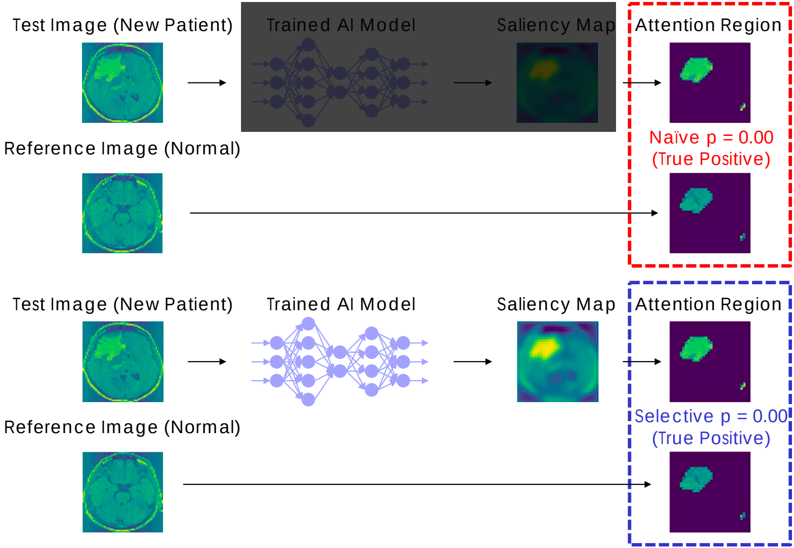
Selection Event by Piecewise-Linear Functions

- A selection event characterized by finite number of piecewise-linear functions looks like:



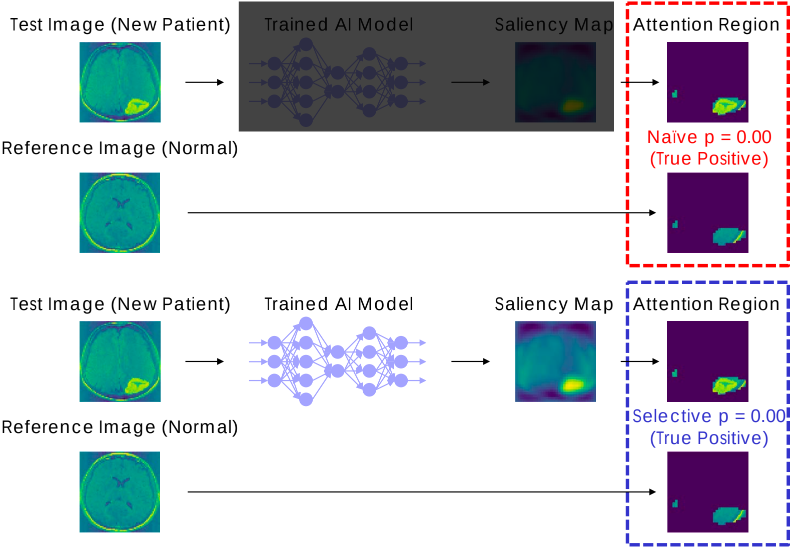
Result (1)

Cases with Real Tumor (Global Null Test / Single Reference)



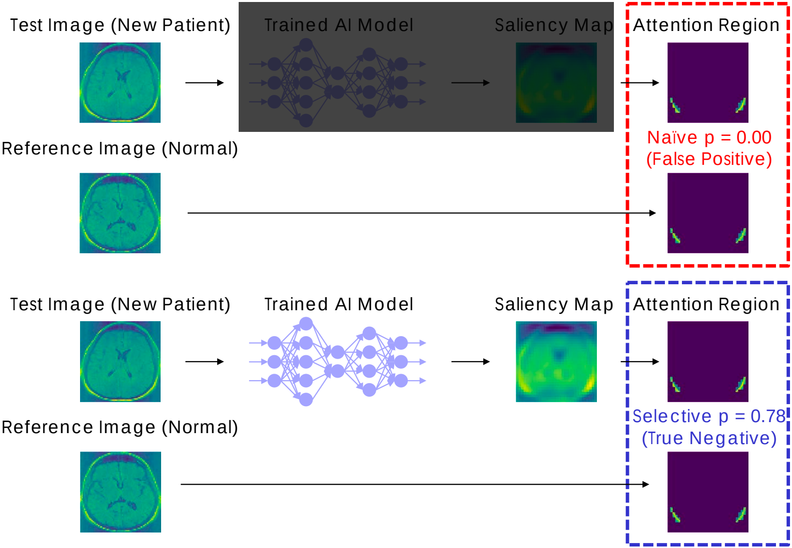
Result (2)

Cases with Real Tumor (Mean Null Test / Single Reference)



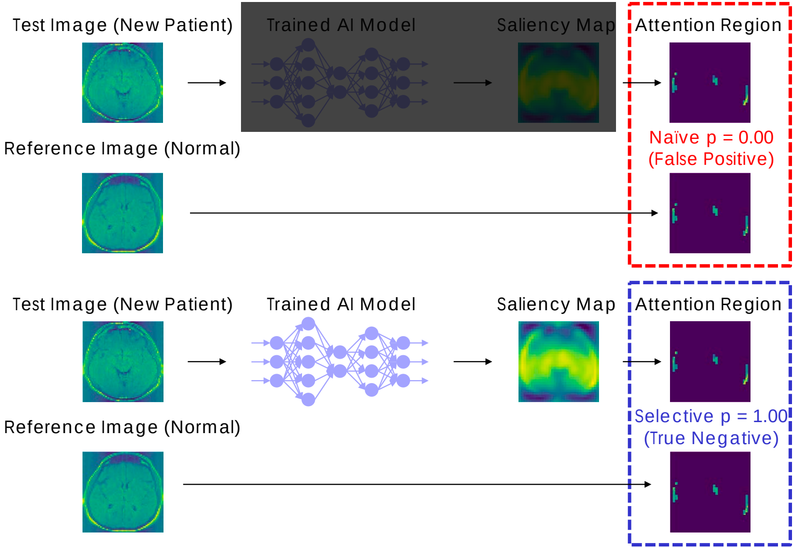
Result (3)

Cases without Real Tumor (Global Null Test)



Result (4)

Cases without Real Tumor (Mean Null Test)



The Messages in This Talk (Revisited)

- Why naive p-values are invalid for AI-driven hypotheses and how we interpret and formulate this issue?

- How selective inference (a new trend in statistics for data-driven hypotheses) resolve this issue?

- How we can compute selective p-values for deep neural network-driven hypotheses?

Summary

- The reliability of data-driven hypotheses cannot be properly evaluated by traditional statistical methods.
- AI-based scientific discovery can be interpreted as a large-scale multiple testing problem where conventional methods cannot be applied.
- Conditional Selective Inference (SI) is a potentially useful tool for correcting the selection bias of data-driven hypotheses.
- Conditional SI is casted into the problem of finding a subset of parametrized datasets that outputs the same hypothesis (inverse problem).
- Conditional SI can be extended to handle an algorithm that can be decomposed into piecewise-linear component.
- Many (seemingly) complicated algorithm (including CNN) can be decomposed into piecewise-linear components, which enables us to employ conditional SI.

Reference I

[1] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor.
Exact post-selection inference, with application to the lasso.
The Annals of Statistics, 44(3):907–927, 2016.

[2] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani.
Selective sequential model selection.
arXiv preprint arXiv:1512.02565, 2015.

[3] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani.
Exact post-selection inference for sequential regression procedures.
Journal of the American Statistical Association, 111(514):600–620, 2016.

[4] Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty.
Selective inference for group-sparse linear models.
In Advances in Neural Information Processing Systems, pages 2469–2477, 2016.

[5] Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi.
Selective inference for sparse high-order interaction models.
In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3338–3347. JMLR. org, 2017.

[6] Shuxiao Chen and Jacob Bien.
Valid inference corrected for outlier removal.
Journal of Computational and Graphical Statistics, pages 1–12, 2019.

[7] Kosuke Tanizaki, Noriaki Hashimoto, Yu Inatsu, Hidekata Hontani, and Ichiro Takeuchi.
Computing valid p-values for image segmentation by selective inference.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9553–9562, 2020.

[8] Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi.
Computing valid p-value for optimal changepoint by selective inference using dynamic programming.
In Advances in Neural Information Processing Systems, pages 11356–11367, 2020.

Reference II

[9] Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi.
Quantifying statistical significance of neural network-based image segmentation by selective inference.
In Advances in Neural Information Processing Systems, 2022.

[10] Xiaoying Tian and Jonathan Taylor.
Selective inference with a randomized response.
The Annals of Statistics, 46(2):679–710, 2018.

[11] Keli Liu, Jelena Markovic, and Robert Tibshirani.
More powerful post-selection inference, with application to the lasso.
arXiv preprint arXiv:1801.09037, 2018.

[12] Kazuya Sugiyama, Vo Nguyen Le Duy, and Ichiro Takeuchi.
More powerful and general selective inference for stepwise feature selection using the homotopy continuation approach.
In Proceedings of the 38th International Conference on Machine Learning, 2021.

[13] Vo Nguyen Le Duy and Ichiro Takeuchi.
Parametric programming approach for more powerful and general lasso selective inference.
In International Conference on Artificial Intelligence and Statistics, pages 901–909. PMLR, 2021.