Explainable Artificial Intelligence (XAI)

Often referred as Interpretable Machine Learning (ML). The terms interpretable and explainable are interchangeably

AI/ML is Important for SDGs Applications: Should We Care About AI/ML Explainability?

Professor Dr. Yasuo MUSASHI, Information Security Division,

Centre for Management and Information Technologies,

Kumamoto University, Japan

Motivations

What Is an Explanation ?

- An explanation is the answer to a **why**-question.*
  - Why did not the treatment work on the patient?
  - Why was my loan or my credit card rejected?
  - Why have we not been contacted by alien life yet?
- AI/ML is important for SDGs: Should we care about AI/ML explainability? YES, Because it is our right !
- YES, Right to an explanation is a right to be given an explanation for an output of the algorithm (See CFPB** and GDPR***).

* Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.[illegible]269. (2017).
** Consumer Financial Protection Bureau: https://www.consumerfinance.gov/
*** General Data Protection Regulation: https://academic.oup.com/idpl/article/7/4/233/4762325

The Right to Explanation is **Mandated**

- Credit scoring in the United States
  - Under the Equal Credit Opportunity Act, creditors are required to notify applicants who are denied credit with specific reasons for the detail.
- European Union
  - The General Data Protection Regulation (**GDPR**) extends the automated decision-making rights in the 1995 Data Protection Directive to provide a legally disputed form of a right to an explanation, stated as such in Recital 71: "[the data subject should have] the right ... to obtain an explanation of the decision reached."
- Explainability in AI and ML is critical. Let’s have a nice discussion!

https://en.wikipedia.org/wiki/Right_to_explanation

Introduction

XAI = To Make AI/ML **Models** and Their **Decisions** Understandable by Human

- AI/ML is great, but **computers usually do not explain their predictions**.
- Moreover, if we don't understand **why they make such a prediction**, can we really trust the prediction?

- For example, AI/ML prediction = today Bitcoin's price will drop!
  - Don't you feel you need to hear computer's explanation **(to understand the model)**, before you sell you Bitcoin?
  - Don't you think you also need to understand why they decide that today Bitcoin's price will drop **(to understand the decision)**, before you sell you Bitcoin?

Model explainability & Decision explainability

- To make the next behavior and predictions of AI/ML systems understandable to humans.*
  - human can understand the cause of a decision.**
  - human can consistently predict the model's result.***
- The higher the explainability or interpretability of a machine learning model, the easier it is for a human to see why specific decisions or predictions have been made.

* https://christophm.github.io/interpretable-ml-book/terminology.html
** Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (20
*** Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretab ormation Processing Systems (2016).

The Importance of XAI

Question: If a machine learning model performs well, why do not we just trust the model and ignore why it made a certain decision?

When We Do **Not Need** XAI ?

- Some models may not require explanations because they are used in a low-risk environment, meaning a mistake will not have serious consequences, (e.g., a movie recommender system)

- The need for interpretability arises when for certain problems or tasks it is not enough to get only the prediction (the **what**).
- The model must also explain how it came to the prediction (the **why**).
  - Example: Is there any racism in a credit approval application, discriminates against a minority?

Understand the **Model** Behavior.

Example: Detect **Edge Cases** for Safety Measure

- Imagine a self-driving car automatically detects cyclists based on a deep learning system.
- You want to be 100% sure that the abstraction the system has learned is error-free, because running over cyclists is quite bad.
- An explanation might reveal that the most important learned feature is to recognize the two wheels of a bicycle, and this explanation helps you think about **edge cases** like bicycles with side bags that partially cover the wheels.

Understand Why that **Decision**, Not the Other. Example: Detect **Bias** in AI/ML Credit Approval Decision.

- AI/ML models pick up **biases** from the <u>training data</u>.
- This can turn your AI/ML into racists, because trained for automatic approval or rejection of credit applications discriminates against a minority that has been historically disenfranchised.
  - The goal is to grant loans only (1) to people who will eventually repay them (granting loans in a low-risk), but also obliged (2) not to discriminate on the basis of certain demographics (compliant way).
  - These may not be covered by the loss function the AI/ML model was optimized for.

XAI Concepts

- Algorithm **Transparency**: How does the algorithm create the model?
  - High transparency (white box) = Algorithms such as the least squares method for linear models are well studied and understood.
  - Less transparent (black box) = Deep learning approaches are less well understood and the inner workings are the focus of ongoing research.
- **Global** Explanation: How does the trained model make predictions? How do parts of the model affect predictions?
- **Local** Explanation: Why did the model make a certain single prediction?

Illustrations of Simple (White Box) Interpretable Models*

With some simple ML algorithm (such as Linear Regression or Logistic Regression), a trained model intuitively easy to be interpreted how it works or how it made a decision.

* Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020. https://christophm.github.io/interpretable-ml-book/

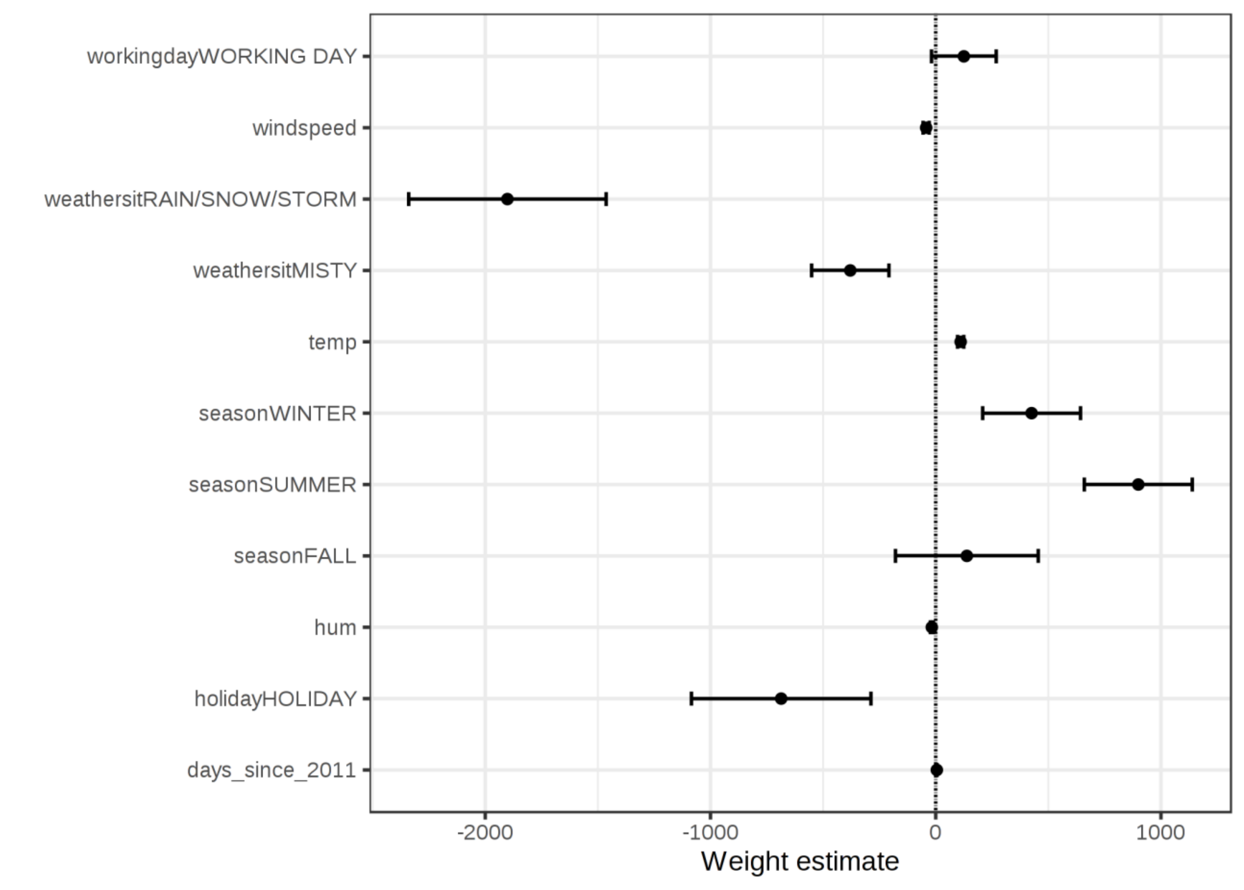
Motivations
- Right to Explanation
- Mandated
Introduction
- Model explainability
- Decision explainability
The Importance
- Not Needed
- Model Behavior
- Single Decision
- Concepts
Interpretable Models
Model-Agnostic
- PDP
- ICE
- Feature Interaction
- LIME
- Anchors
- Shapley Values
- SHAP
TODO: Our Journal Paper

Interpreting a **Linear Regression** Model

to predict the number of rented bikes on a particular day, given weather and calendar information.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- A linear regression model predicts the target as a weighted sum of the feature inputs.
- The linearity of the learned relationship makes the interpretation easy.
- The weight plot shows that rainy/snowy/stormy weather has a strong negative effect on the predicted number of bikes.



Interpreting a **Logistic Regression** Model

to predict cervical cancer based on some risk factors.

$$P(y^{(i)}=1)=\frac{1}{1+exp(-(\beta_0+\beta_1x_1^{(i)}+\ldots+\beta_px_p^{(i)}))}$$

- An increase in the number of diagnosed STDs (sexually transmitted diseases) changes (increases) the odds of cancer vs. no cancer by a factor of 2.27, when all other features remain the same.
- For women using hormonal contraceptives, the odds for cancer vs. no cancer are by a factor of 0.89 lower, compared to women without hormonal contraceptives, given all other features stay the same.

	Weight	Odds ratio	Std. Error
Intercept	-2.91	0.05	0.32
Hormonal contraceptives y/n	-0.12	0.89	0.30
Smokes y/n	0.26	1.30	0.37
Num. of pregnancies	0.04	1.04	0.10
Num. of diagnosed STDs	0.82	2.27	0.33
Intrauterine device y/n	0.62	1.86	0.40

Model-Agnostic Methods for Complex (Black Box) AI/ML Model

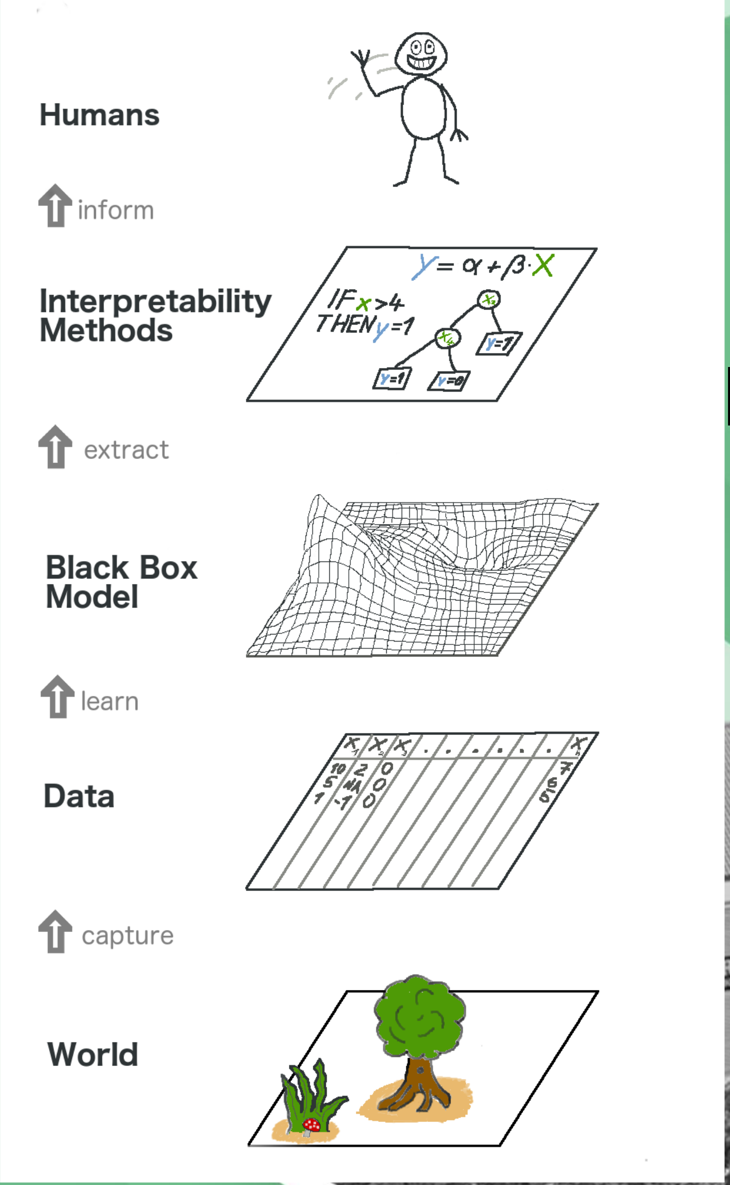
Separating the explanations from the ML model.

ISICO 2021

Model-Agnostic

熊本大学 Kumamoto University

- Model flexibility
- The interpretation method can work with any machine learning model, such as random forests and deep neural networks.

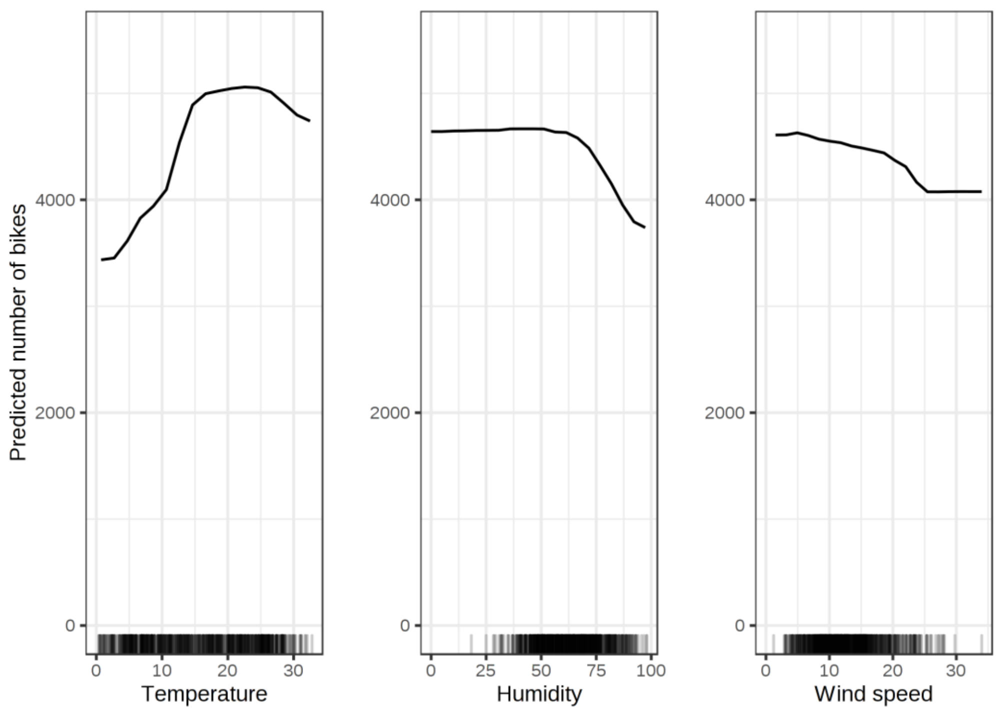


Source: https://christophm.github.io/interpretable-ml-book/

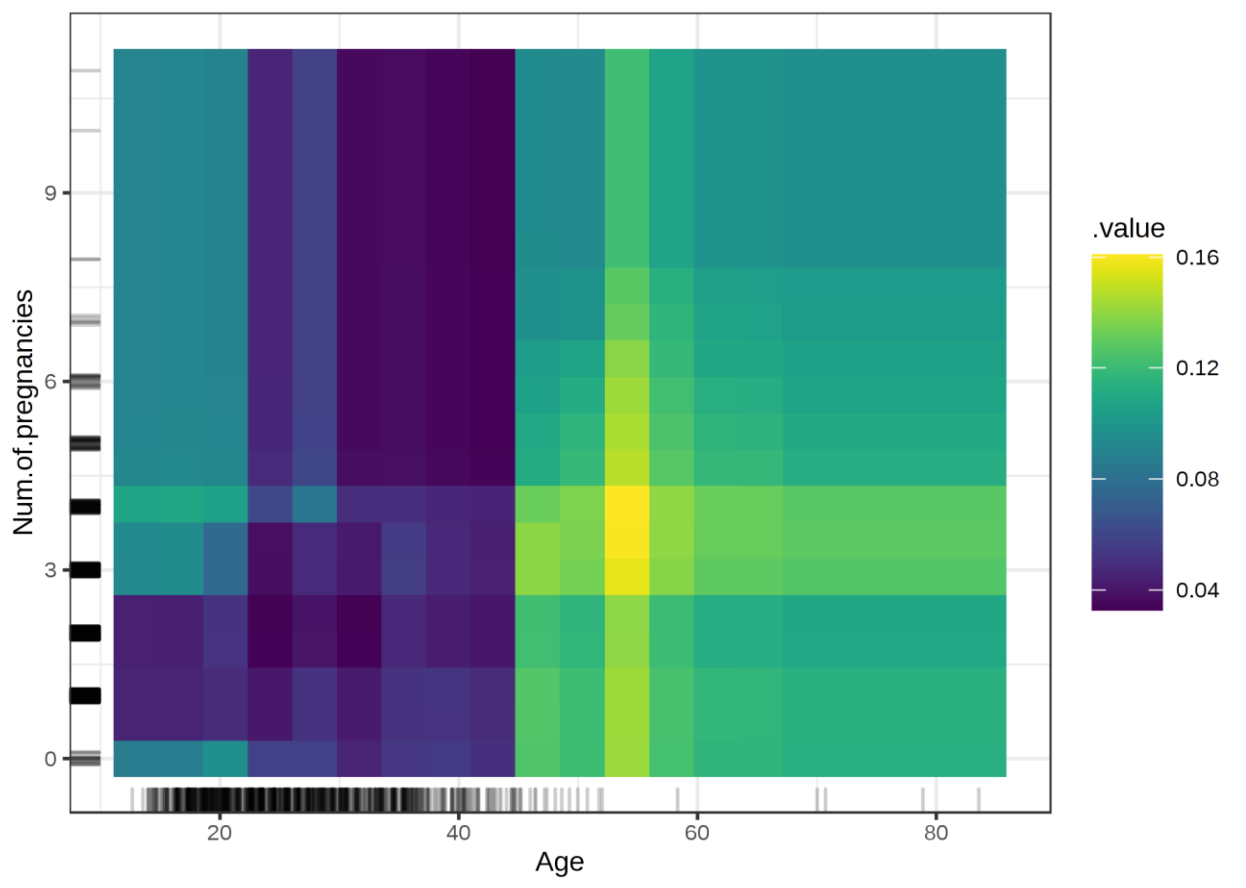
Partial Dependence Plot (PDP)

shows the marginal effect one or two features have on the predicted outcome of a machine learning model

The hotter, the more bikes are rented.



cancer probability increase at age 45. For ages below 25, women who had 1 or 2 pregnancies have a lower predicted cancer risk

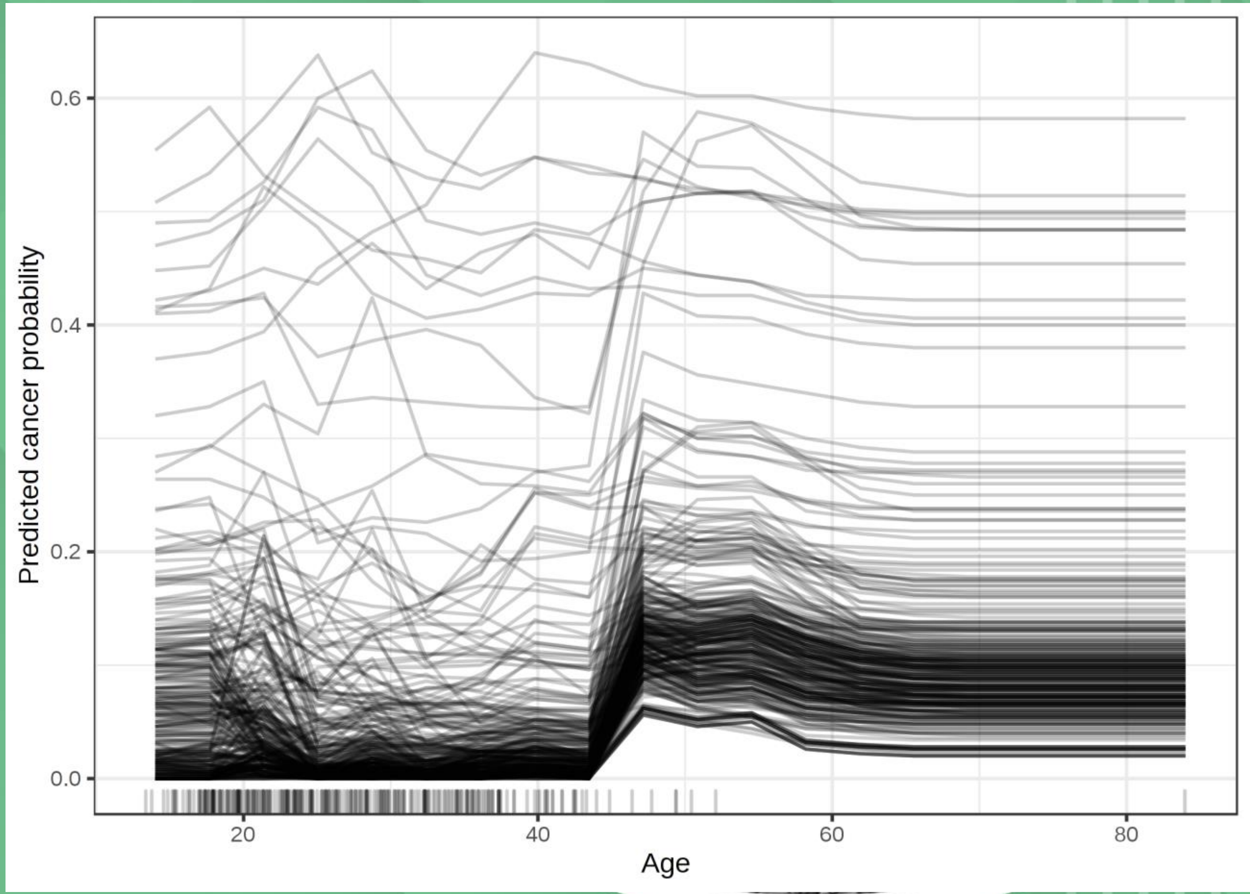


Source: https://christophm.github.io/interpretable-ml-book/

Individual Conditional Expectation (ICE)

shows how the instance's prediction changes when a feature does

- Probability of cancer, when age changes:
- For most women there is an increase in predicted cancer probability with increasing age.
- For some women with a predicted cancer probability above 0.4, the prediction does not change much at higher age.

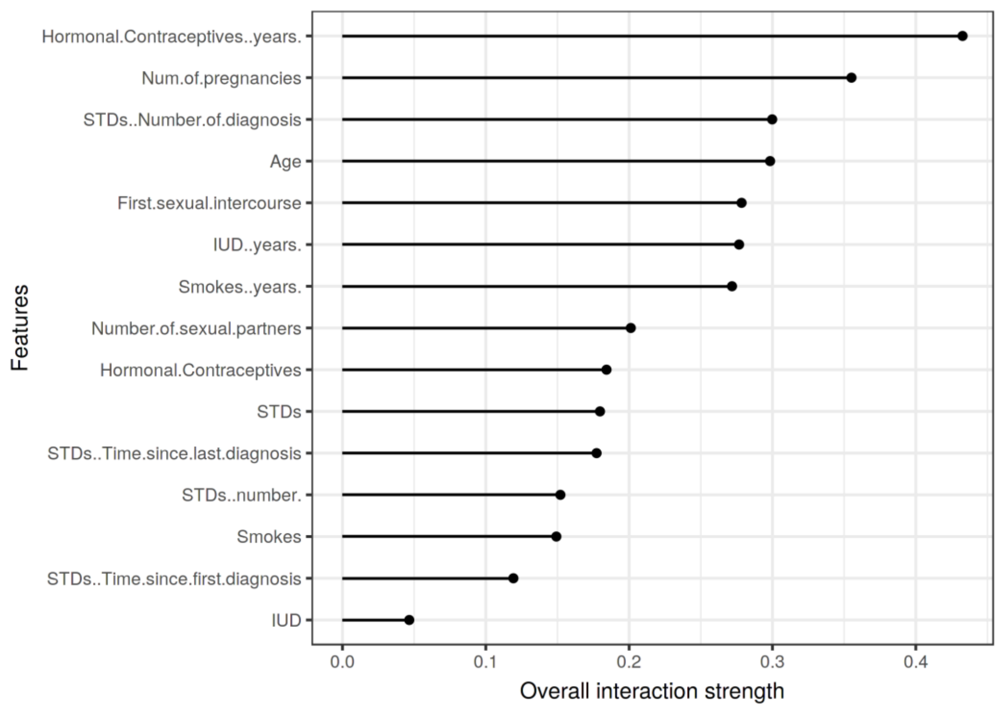


Source: https://christophm.github.io/interpretable-ml-book/

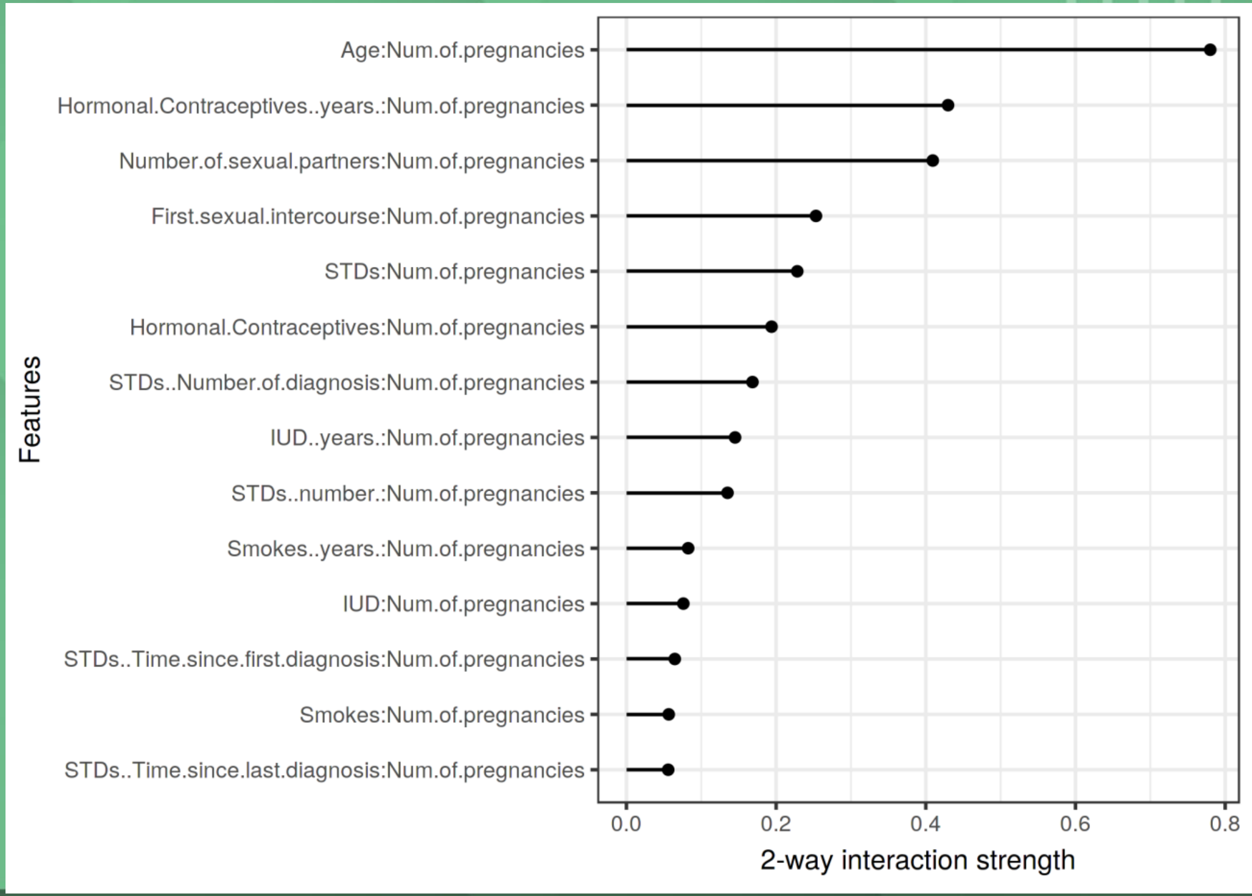
Feature Interaction

When features interact with each other, the effect of one feature depends on the value of the other feature

The years on hormonal contraceptives has the highest relative interaction effect with all other features



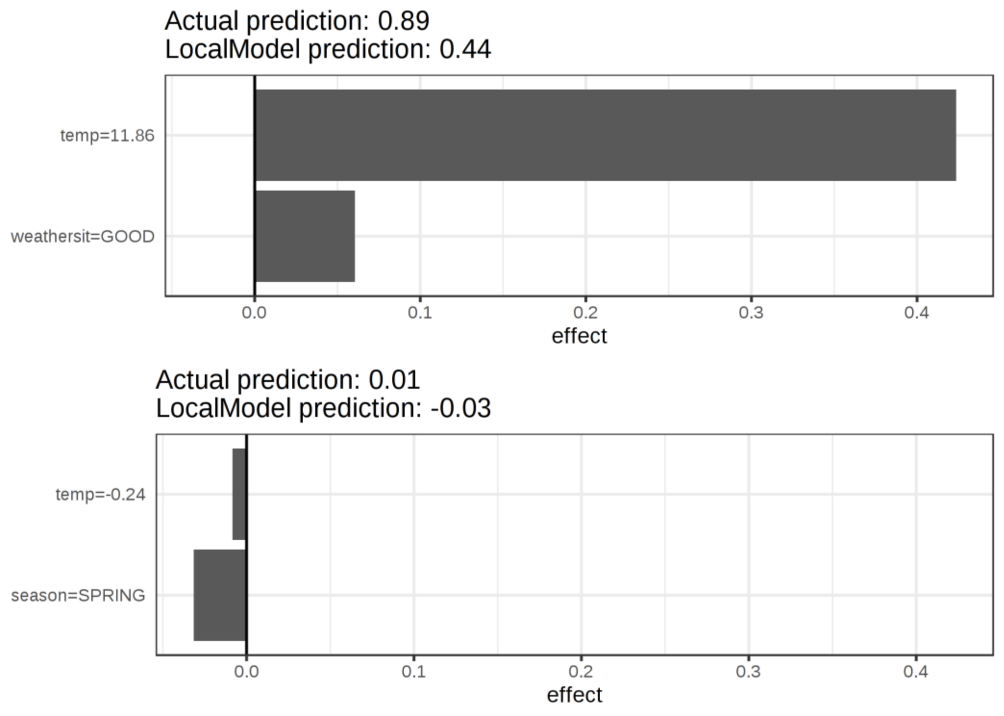
There is a strong interaction between the number of pregnancies and the age



Source: https://christophm.github.io/interpretable-ml-book/

Local interpretable model-agnostic explanations (**LIME**) used to explain individual predictions of black box machine learning models.

Warmer temperature and good weather situation have a positive effect on the prediction



Left: Image of a bowl of bread. Middle and right: LIME explanations for the top 2 classes (bagel, strawberry) for image classification

Source: https://christophm.github.io/interpretable-ml-book/

Scoped Rules (**Anchors**) explains individual predictions of any black-box classification model by finding a decision rule that "anchors" the prediction sufficiently.

Whether or not a passenger survives the Titanic disaster. One exemplary individual and the model's prediction

Feature	Value
Age	20
Sex	female
Class	first
TicketPrice	300$
More attributes	...
Survived	**true**

And the corresponding anchors explanation is:

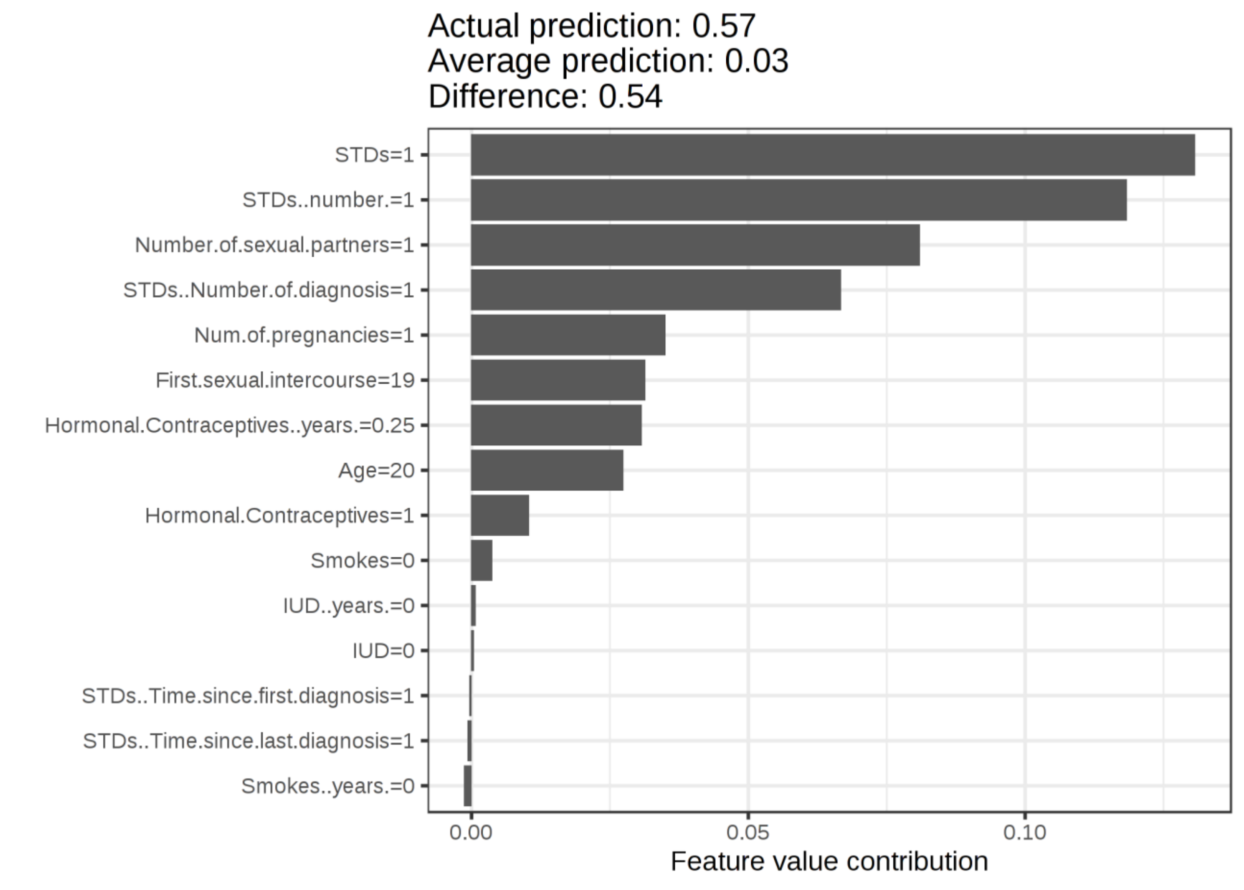
IF `SEX = female`
AND `Class = first`
THEN PREDICT `Survived = true`
WITH PRECISION `97%`
AND COVERAGE `15%`

Source: https://christophm.github.io/interpretable-ml-book/

Shapley Values

-- a method from coalitional game theory -- tells us how to fairly distribute the "payout" among the features.

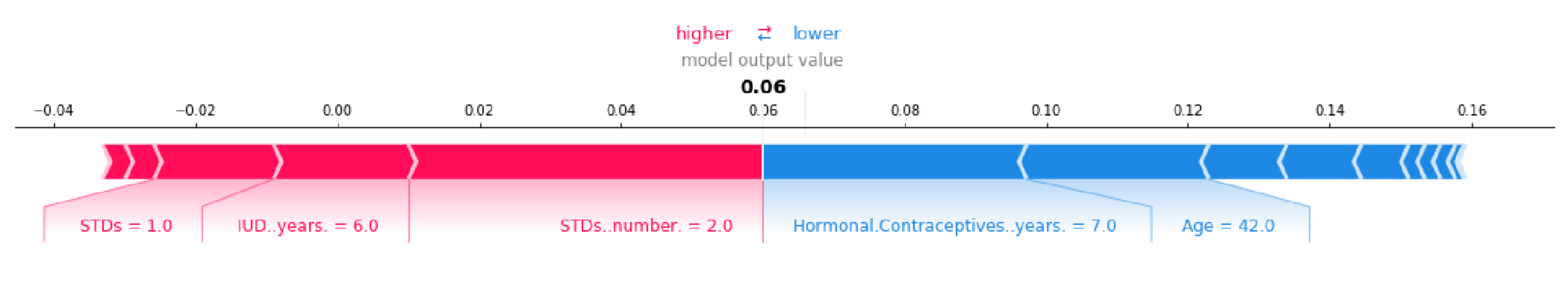
- With a prediction of 0.57, this woman's cancer probability is 0.54 above the average prediction of 0.03.
- The number of diagnosed STDs increased the probability the most.
- The sum of contributions yields the difference between actual and average prediction (0.54).



SHAP (SHapley Additive exPlanations)

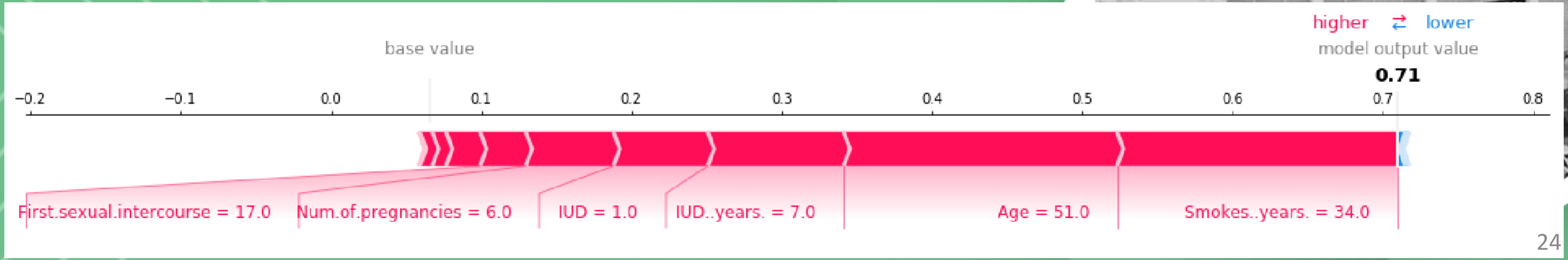
explain individual predictions based on the game theoretically optimal Shapley Values.

- The first woman has a low predicted risk of 0.06.



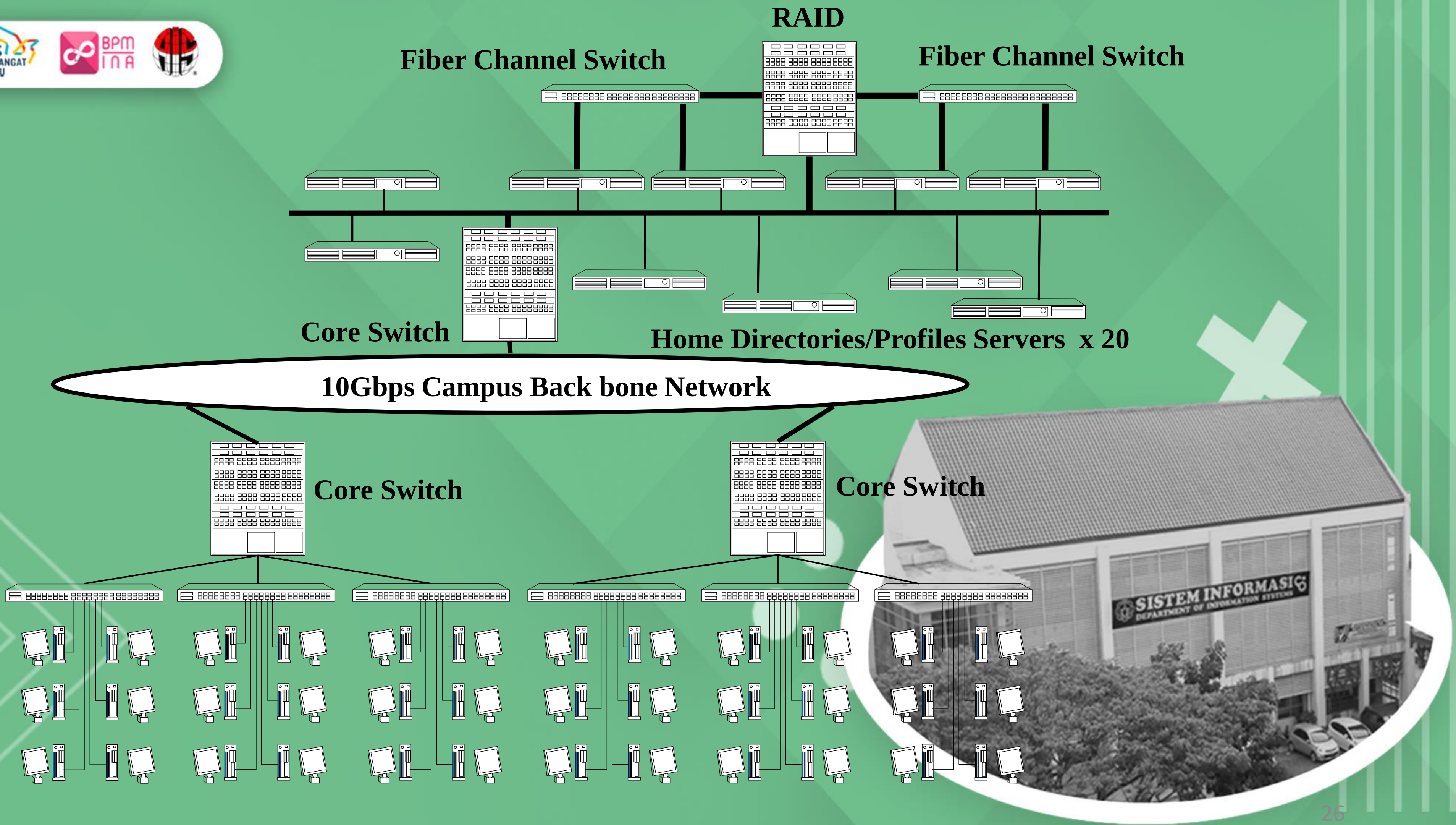
age.

- The second woman has a high predicted risk of 0.71.

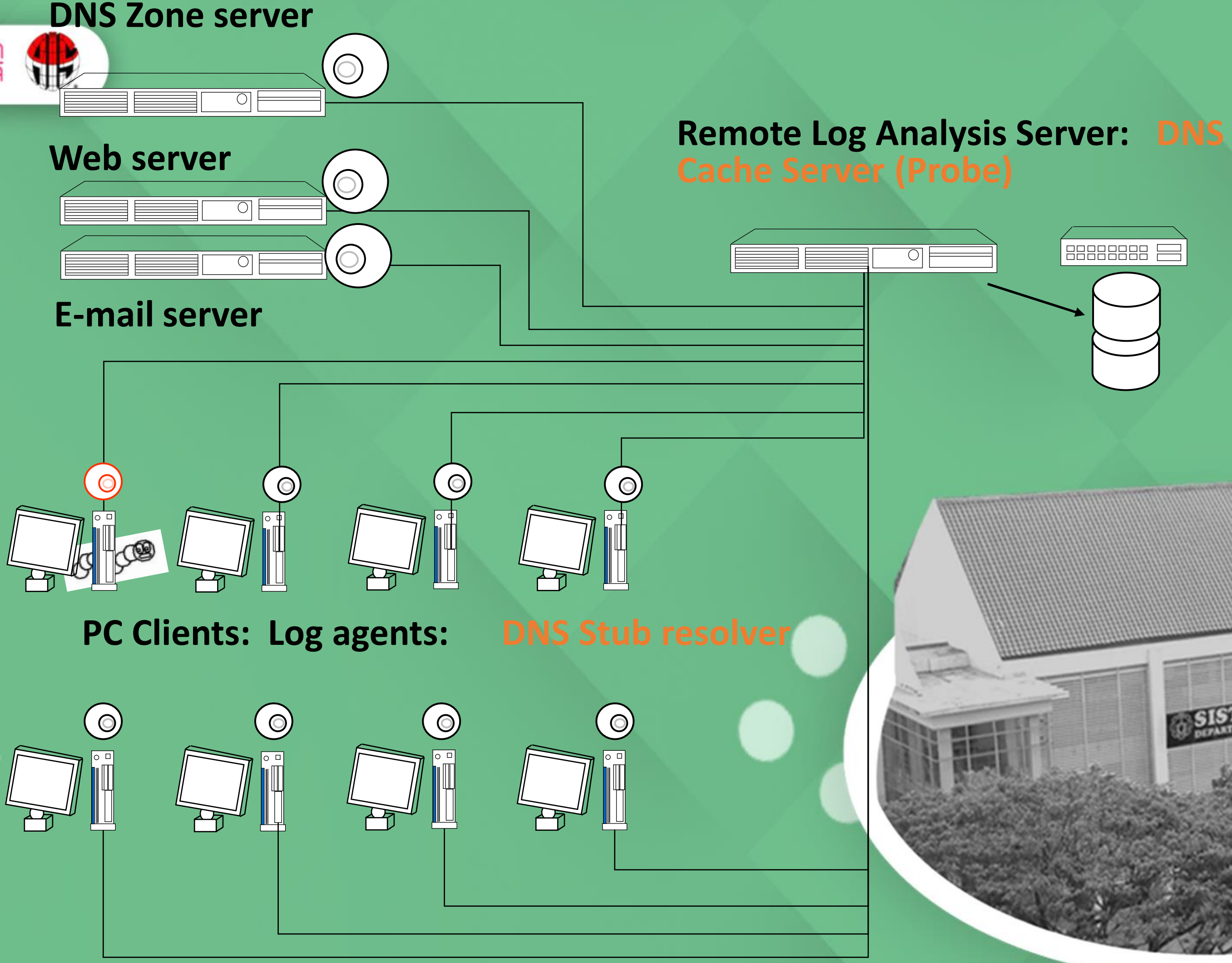


[redacted]

A Campus Network Systems

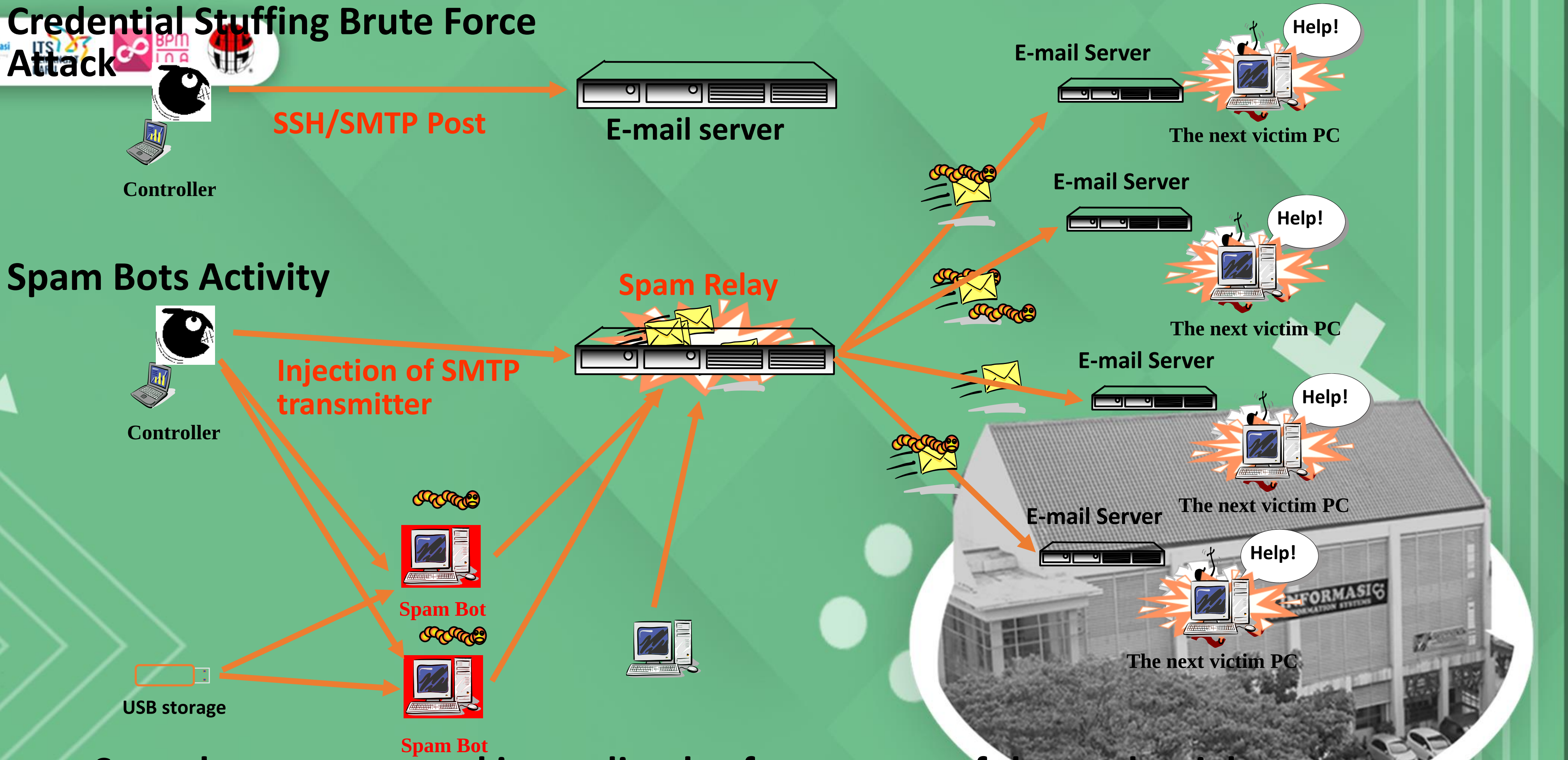


DNS as a Log Agent Service: DNSaaLAS



Motivations
- Right to Explanation
- Mandated
Introduction
- Model explainability
- Decision explainability
The Importance
- Not Needed
- Model Behavior
- Single Decision
- Concepts
Interpretable Models
Model-Agnostic
- PDP
- ICE
- Feature Interaction
- LIME
- Anchors
- Shapley Values
- SHAP
TODO: Our Journal Paper

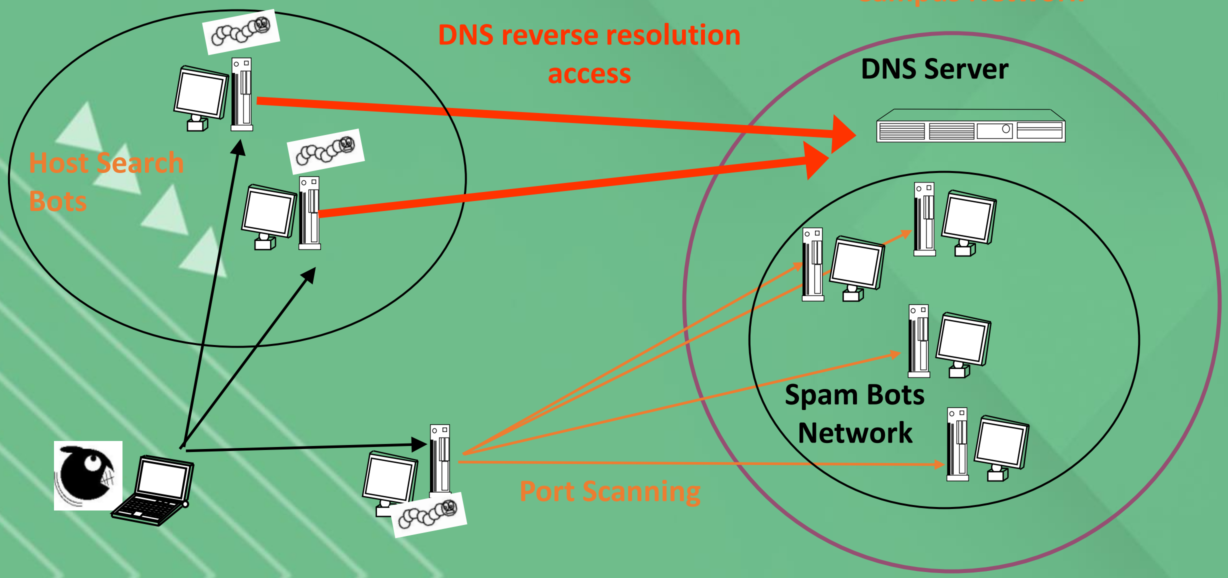
Credential Stuffing Brute Force Attack and Spam Relay



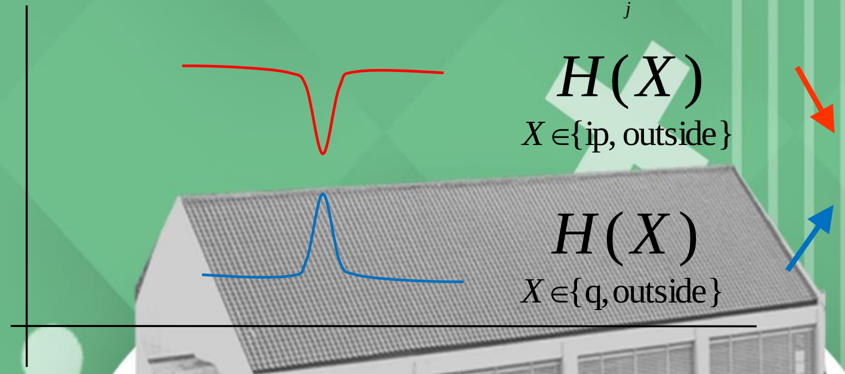
Spam bots gets started immediately after success of the credential stuffing brute force attack.

Host Search: FQDN harvesting

- **DNS Host Search Attack like Directory Harvesting/Service Attack**
- **DNS reverse resolution access from the Internet sites**



$$H(X) = -\sum_{i \in X} P(i) \log_2 P(i) \qquad P(i) = \frac{freq(i)}{\sum_j freq(j)}$$



A Host Search activity model

```
Oct 12 08:38:24 kun named[533]: client 133.95.xxx.yyy#39815: query: 130.13.194.xxx.in-addr.arpa IN PTR
Oct 12 08:38:25 kun named[533]: client 133.95.xxx.yyy#39825: query: dmea.net IN MX
Oct 12 08:38:43 kun named[533]: client 133.95.xxx.yyy#40010: query: mxwall03.hkabc.net IN A
```

Motivations
- Right to Explanation
- Mandated

Introduction
- Model explainability
- Decision explainability

The Importance
- Not Needed
- Model Behavior
- Single Decision
- Concepts

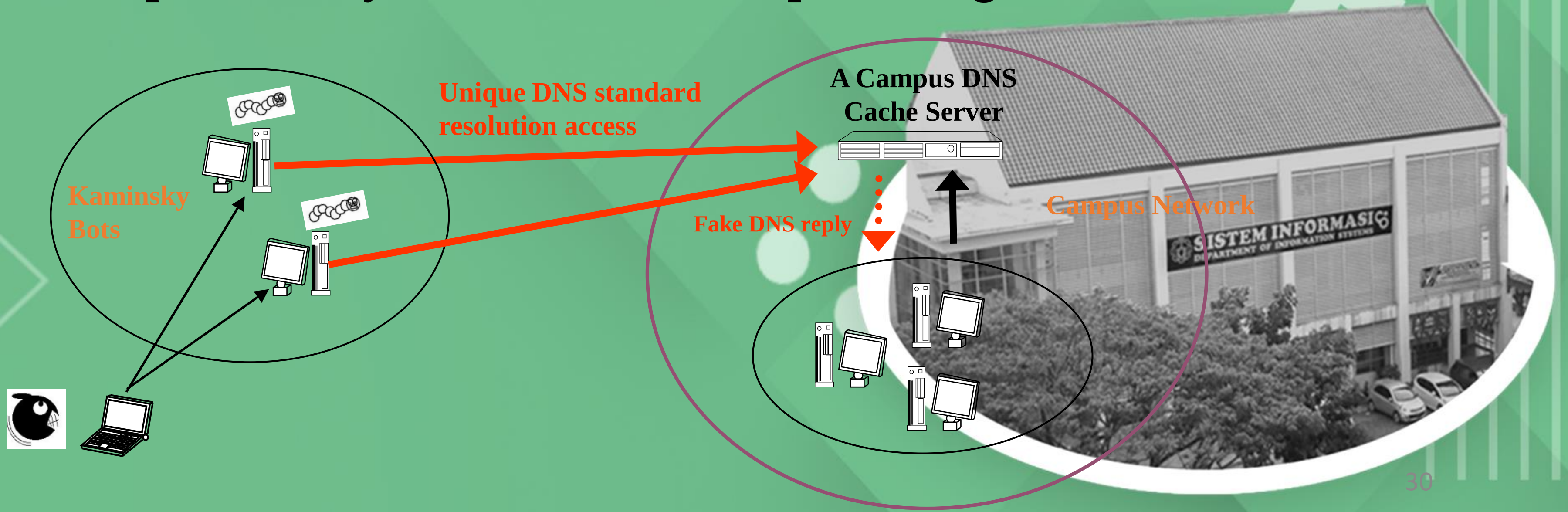
Interpretable Models

Model-Agnostic
- PDP
- ICE
- Feature Interaction
- LIME
- Anchors
- Shapley Values
- SHAP

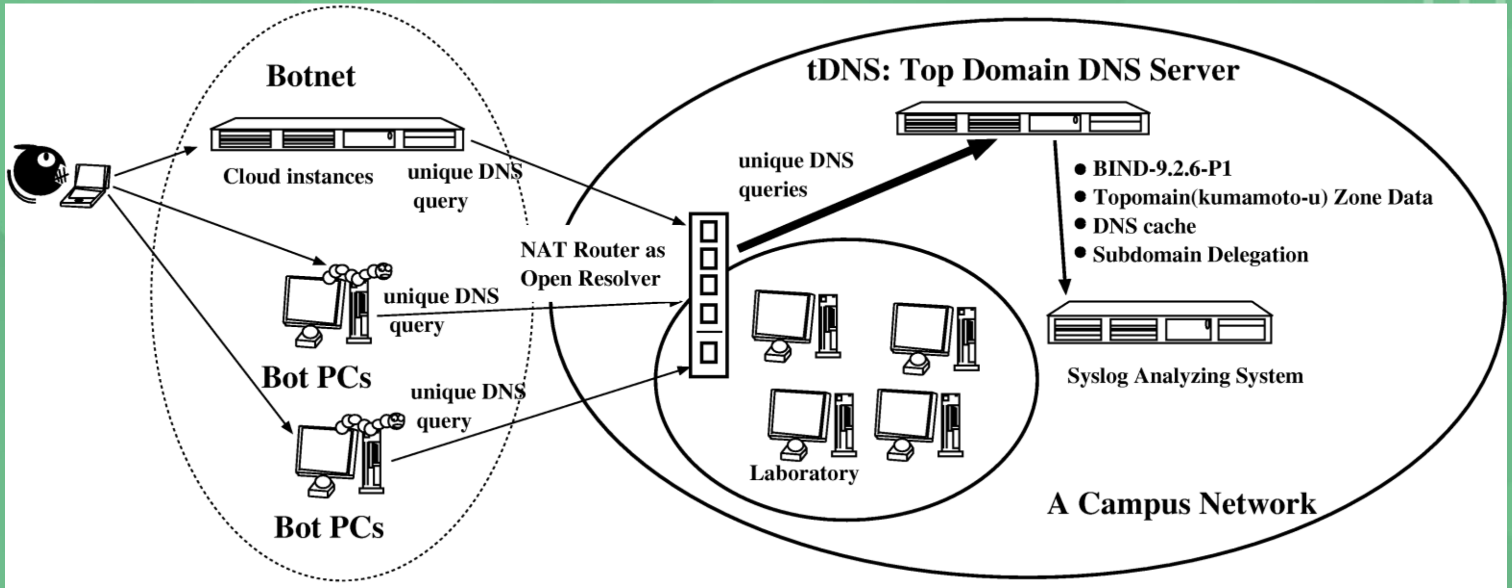
TODO: Our Journal Paper

Kaminsky DNS Cache Poisoning Attack Model: DDoS

- It sends a lot of unique DNS query requests and their query replies to a DNS cache server
- To generate many recursive DNS query requests to upper sites
- In other word, to generate many DNS replies from the upper sites
- To raise the probability in the DNS cache poisoning attack

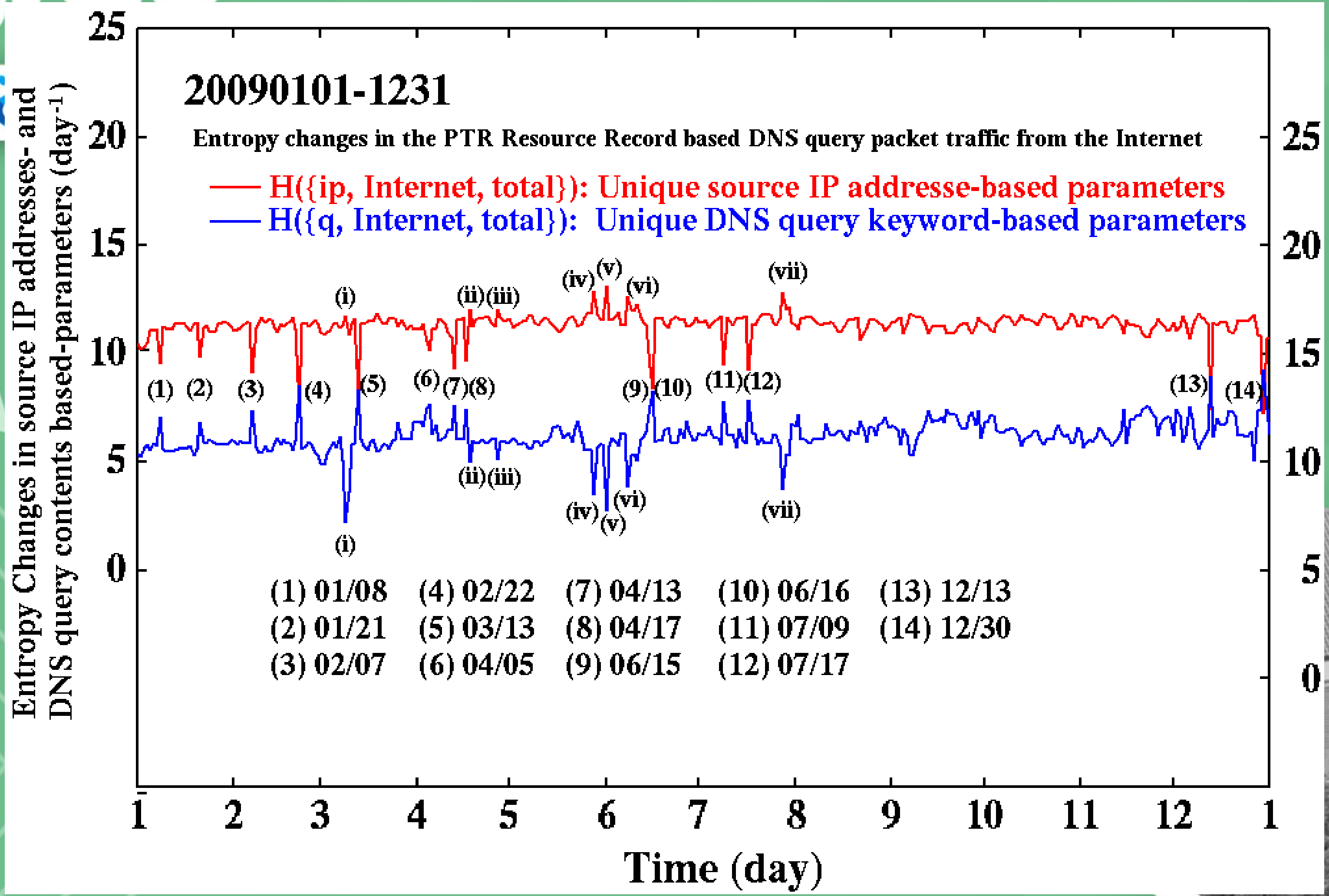


DDoS Attack: Open DNS Resolver



- A Component for Reflection DDoS (Distributed Denial of Service) attack
  - 100Gbps in Sprint Networks
- Kanminsky type DNS cache poisoning attack but no DNS reply packets
- Water Torture: A Slow Drip DNS DDoS Attack https://blog.secure64.com/?p=377

DNS Entropy Changes in the Reverse Name Resolution



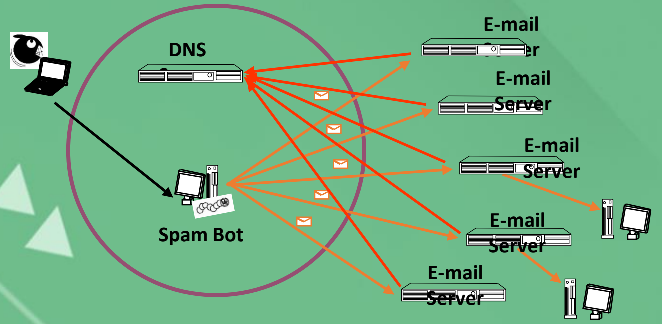
$H(X)_{X\in\{ip,\,outside\}} \searrow$, $H(X)_{X\in\{q,\,outside\}} \nearrow$ { (1)-(14)}

$H(X)_{X\in\{ip,\,outside\}} \nearrow$, $H(X)_{X\in\{q,\,outside\}} \searrow$ {(i)-(vii)}

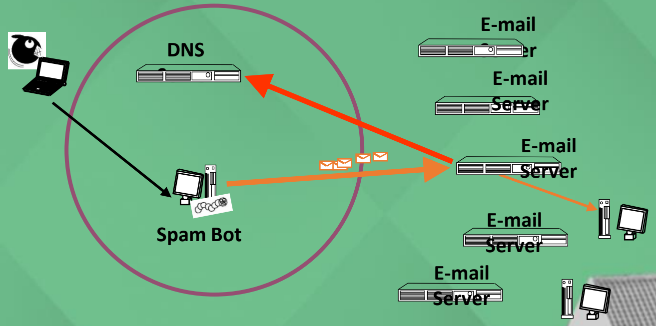
Totally, 18 significant peaks can be observed, consisting of 14 and 8 peaks for HS and RA activities, but no TA activity can be shown.

Random Attack and Target Attack

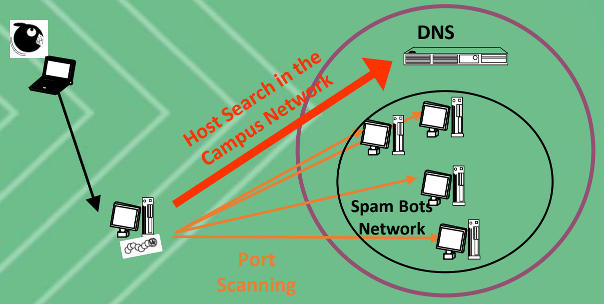
Random Spam Bots (RSB)



Targeted Spam Bots (TSB)



Host Search Attack



D. A. Ludeña Romaña, S. Kubota, K. Sugitani, and Y. Musashi, *IPSJ SIG Technical Reports, the 1st Internet and Operational Technologies (IOT01)*, Vol. **2008**, No.37, pp.103-108 (2008).

A Campus Network Systems

A Campus Network Systems

A Campus Network Systems

A Campus Network Systems

A Campus Network Systems

ISICO 2021

熊本大学 Kumamoto University

A Campus Network Systems

A Campus Network Systems