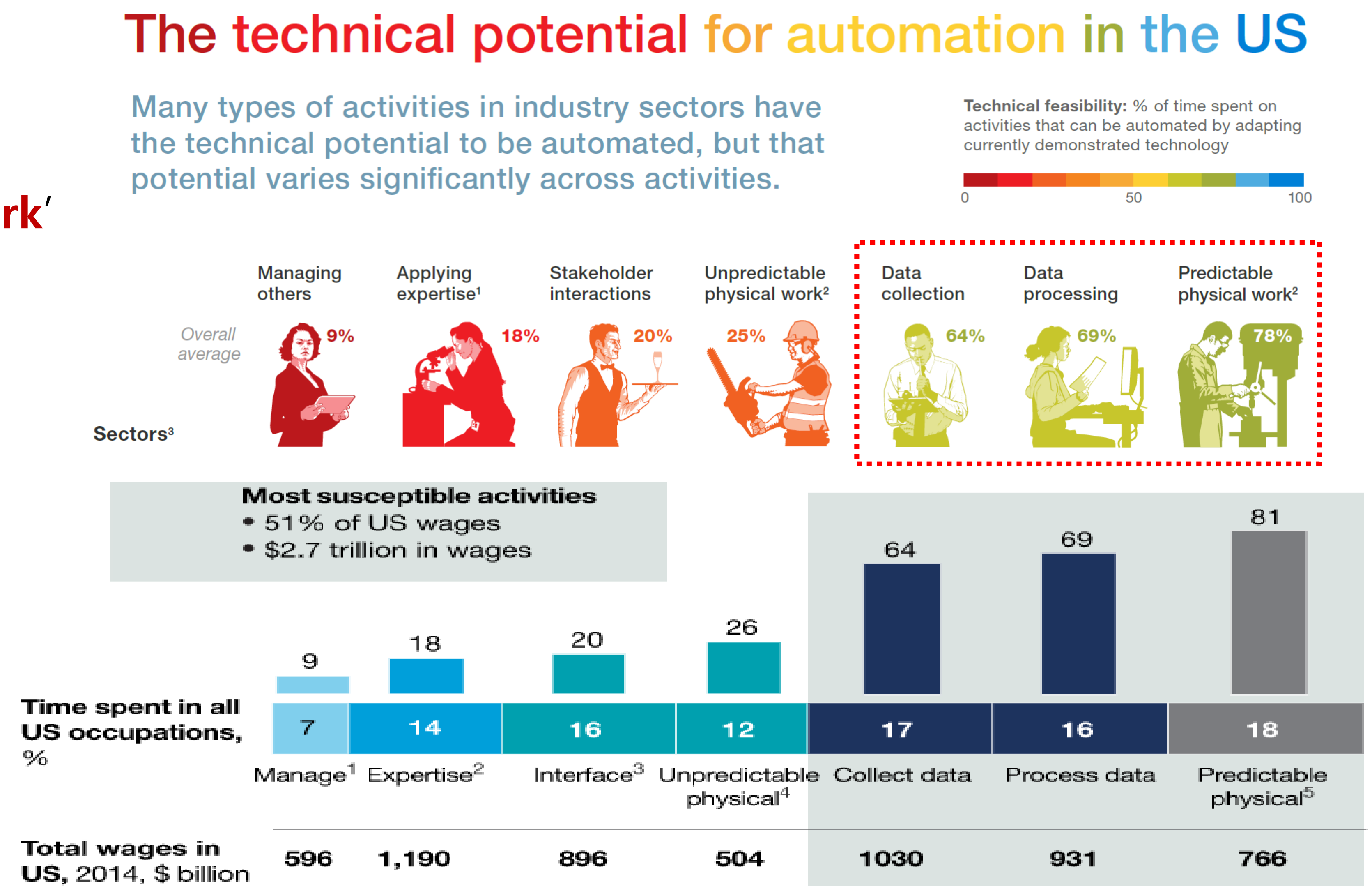
Conditions for Trustworthy AI:
Explainable Artificial Intelligence

Jaesik Choi

Explainable Artificial Intelligence Center
Graduate School of Artificial Intelligence
KAIST

In 2025,
estimated economic impact of
‘**Automation of Knowledge work**’
may reach up to
6.7 trillion US dollar.

In US,
51% of US wages or
$2.7 trillion in wages
could be automated.



Automation of Knowledge Work [McKinsey 2013]

DARPA Grand Challenge 2005

Say Hello to Waymo 2016

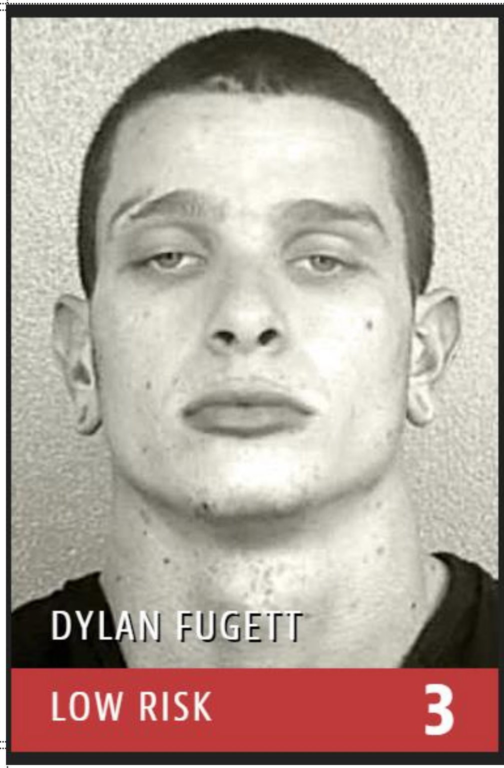
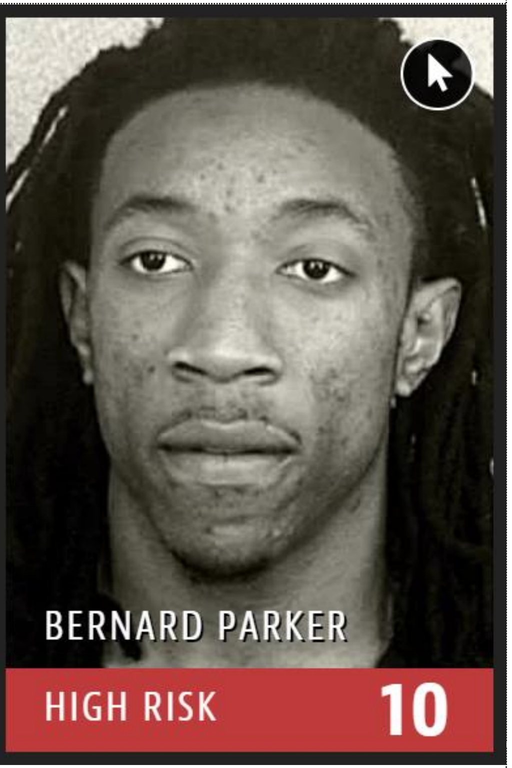
Uber's first car accident - [Death of Elaine Herzberg]

Many, complex AI systems are not transparent to see the mechanisms inside!

Uber's self-driving car killed a pedestrian (Marc 18th, 2018)
The 'safety driver' was watching a TV show (June 22th, 2018)

Do We Understand AI Systems Enough?

COMPAS: Prediction of Crime

<table>
<tr><td>Prior Offense</td><td>1 attempted burglary</td><td>1 resisting arrest without violence</td></tr>
<tr><td>COMPAS' decision</td><td>
</td><td>
</td></tr>
<tr><td>Subsequent Offenses</td><td>3 drug possessions</td><td>None</td></tr>
</table>

AI algorithms are exposed to

(1) data bias,
(2) model bias, and
(3) algorithmic bias

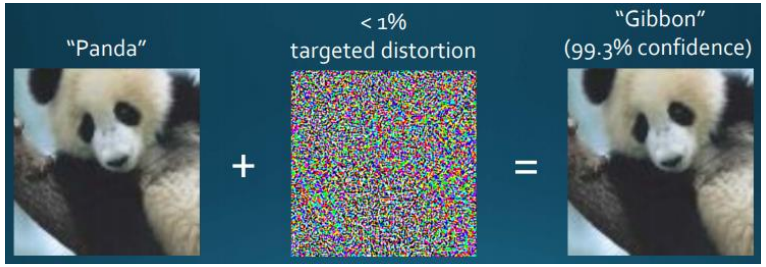
Do We Understand AI Systems Enough?

Article	Contents
13-14. **Right to explanation**	A data subject has the right to "**meaningful information about the logic involved**" when decision is made automatically.
EU administration	When violated **4% of global revenue** will be fined.
Enact	**May 28th, 2018**

EU General Data Protection Regulation (GDPR)

DESCRIBE	CATEGORIZE	EXPLAIN
Handcrafted Knowledge	*Statistical Learning*	*Contextual Adaptation*

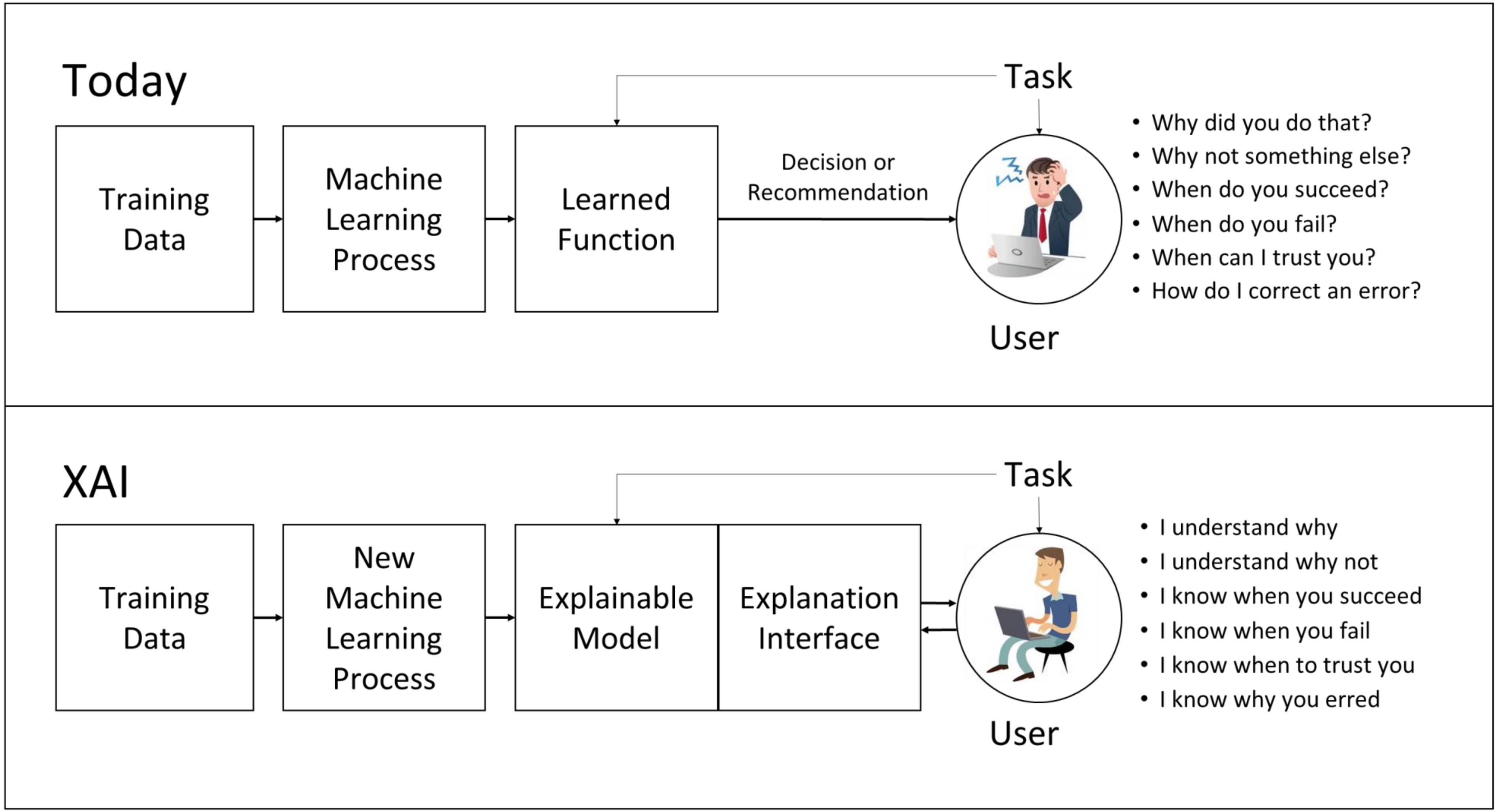
Statistically impressive, but individually unreliable



Inherent flaws can be exploited

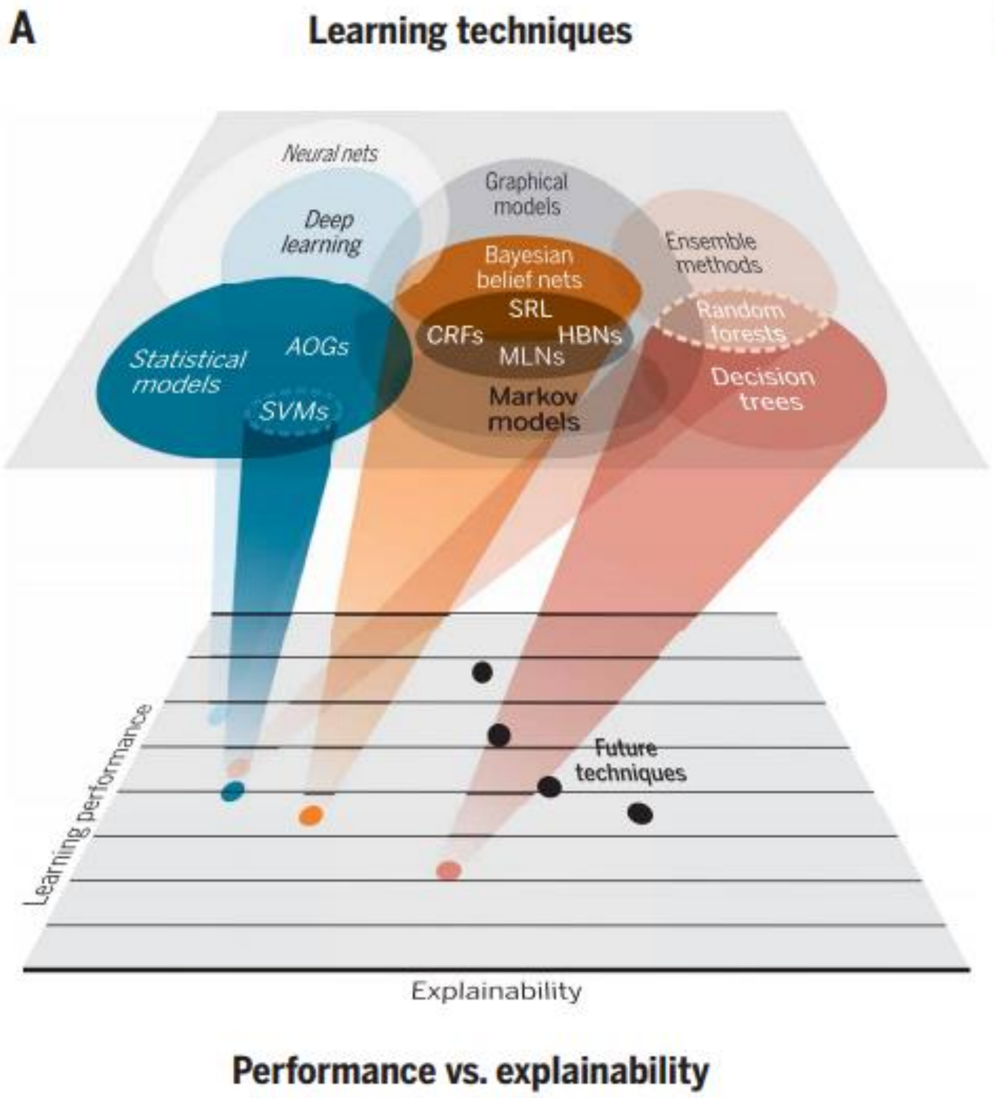
Skewed training data creates Maladaptation

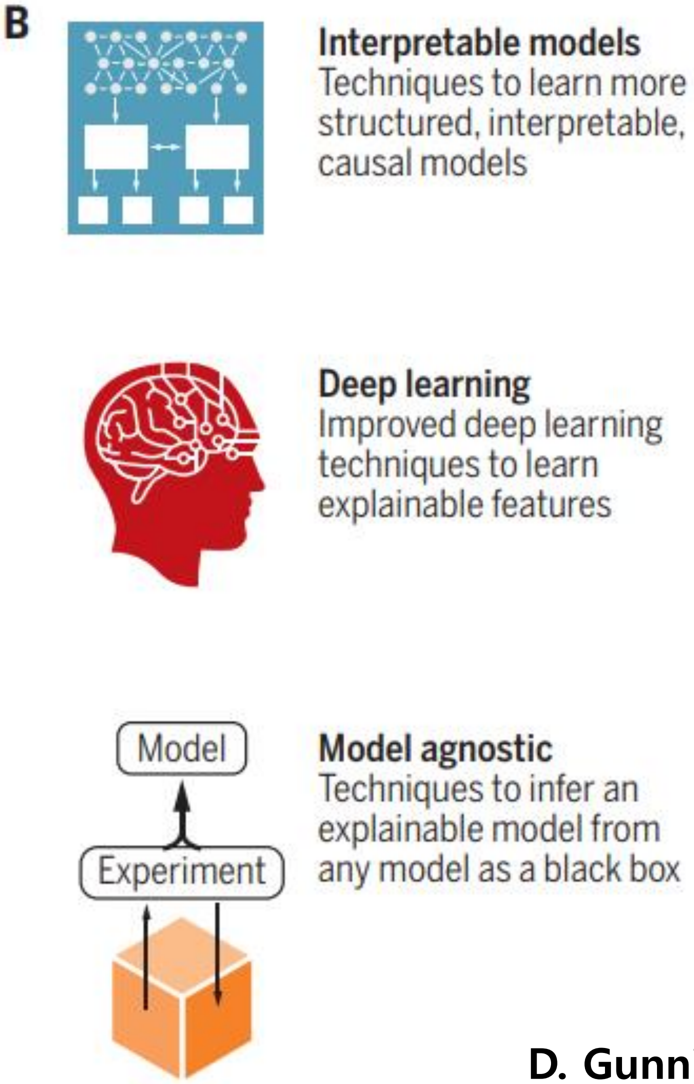
A DARPA Perspective on AI – Three Waves of AI



DARPA

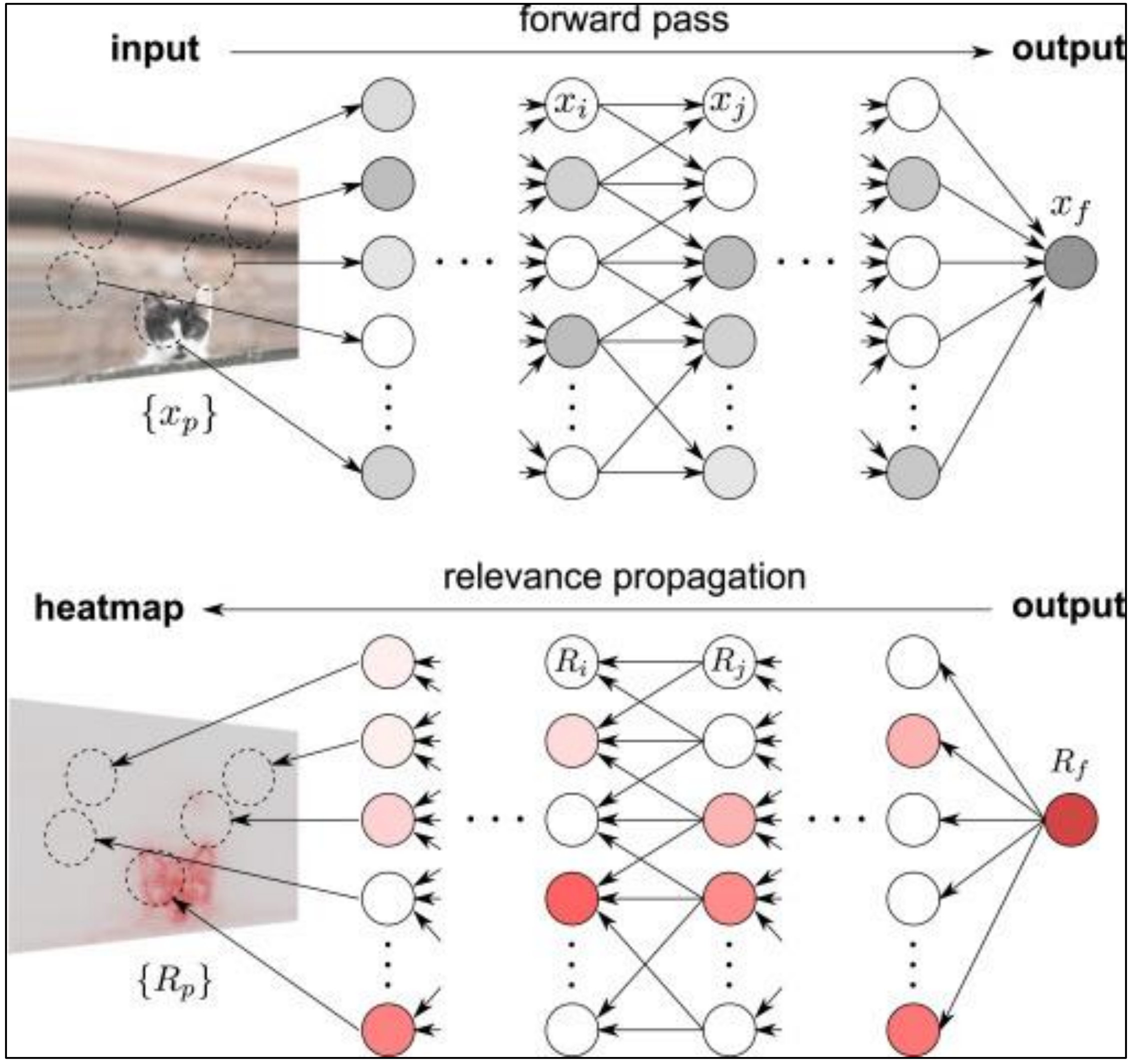
Explainable AI – Performance vs. Explainability





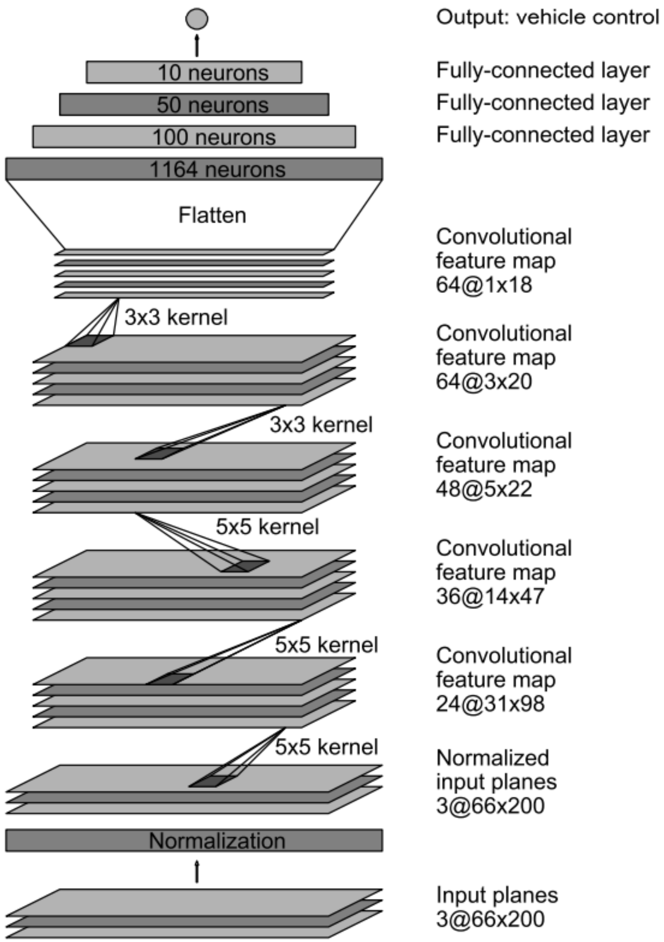
D. Gunning et al., Science Robotics, 2019

Explainable AI – Performance vs. Explainability



[Image courtesy of Klaus Muller]

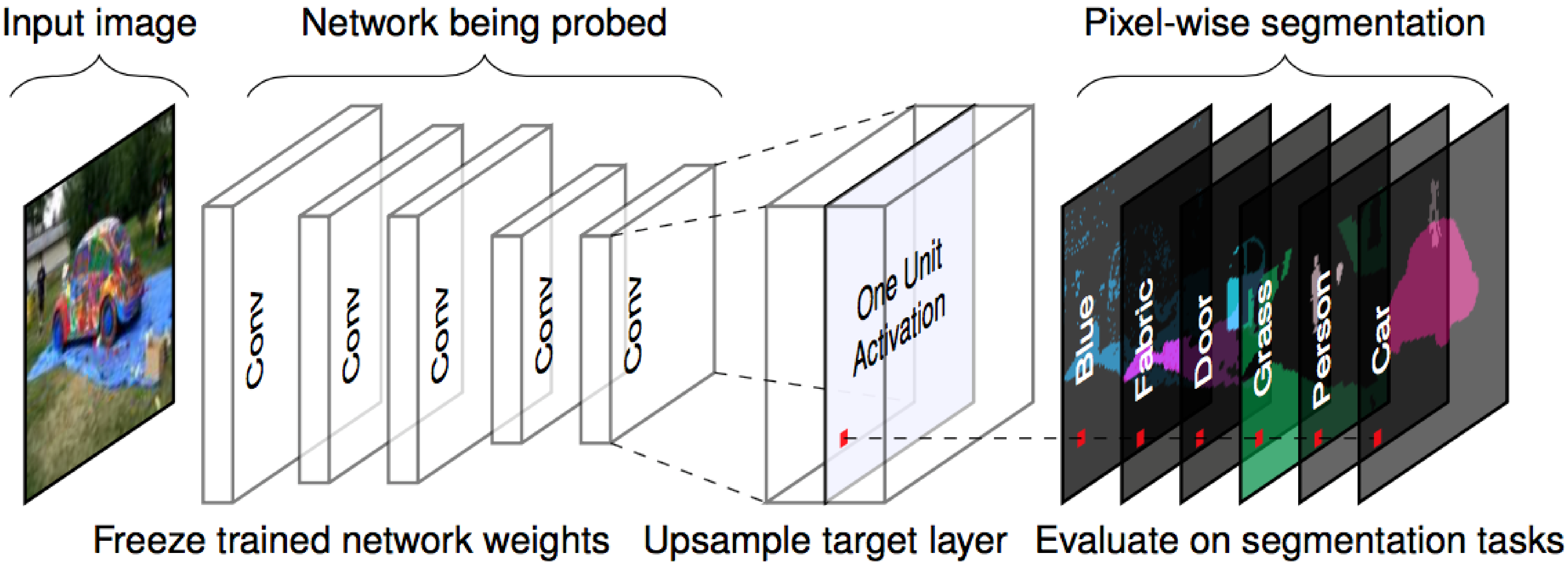
Input Attribution Methods



PilotNet Architecture, NVIDIA/Google, 2017

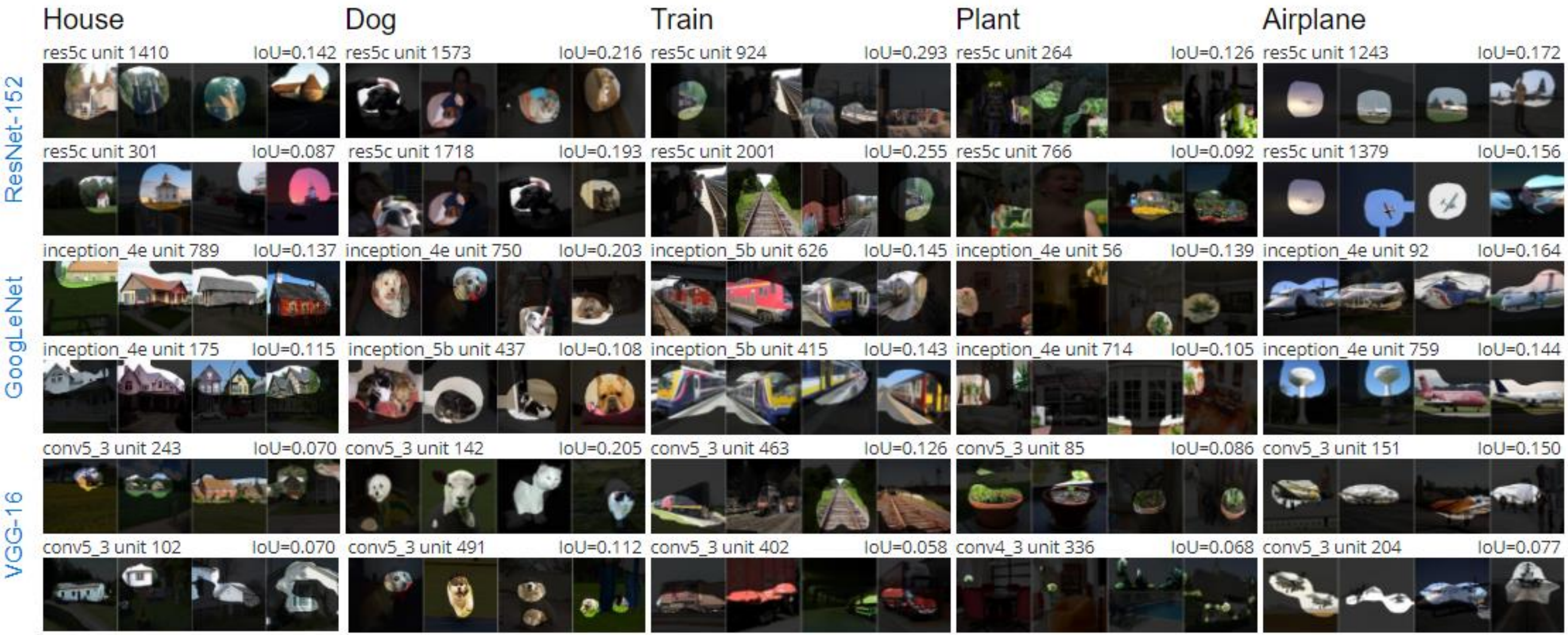
Input Attributions of PilotNet

Explaining Decision of Autonomous Driving



Network Dissection

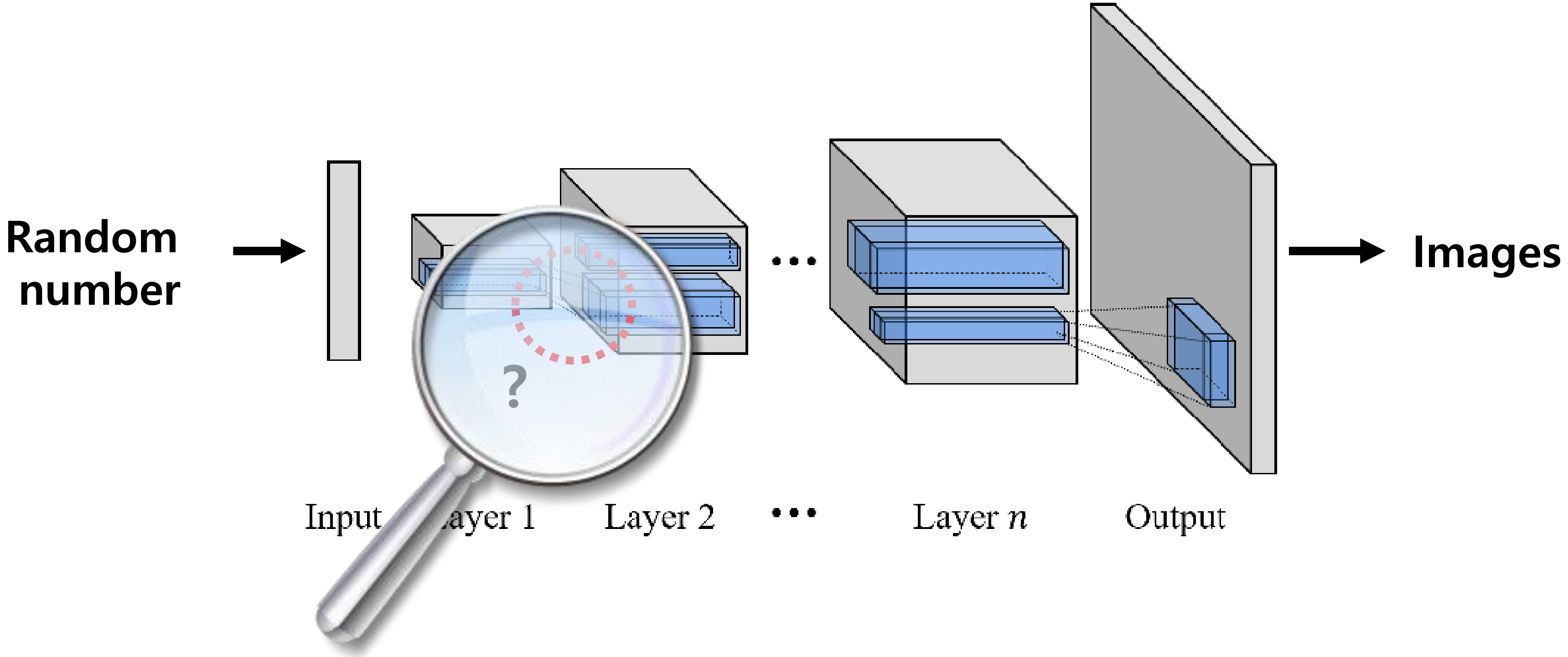
D. Bau et. al., CVPR, 2017
D. Bau et. al., PNAS, 2020



Network Dissection

D. Bau et. al., CVPR, 2017
D. Bau et. al., PNAS, 2020

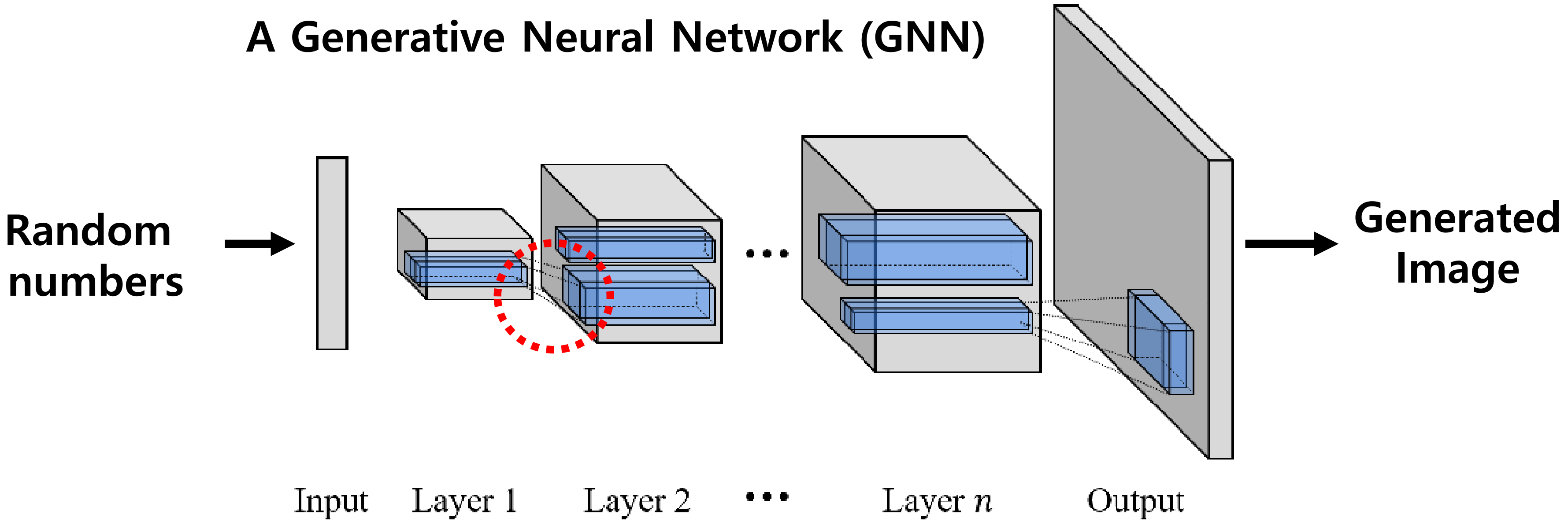
A Generative Neural Network (GNN)

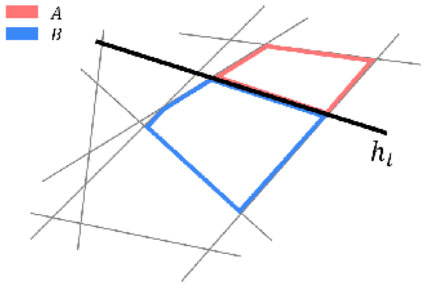


Explorative Generative Boundary Aware Sampling (E-GBAS)

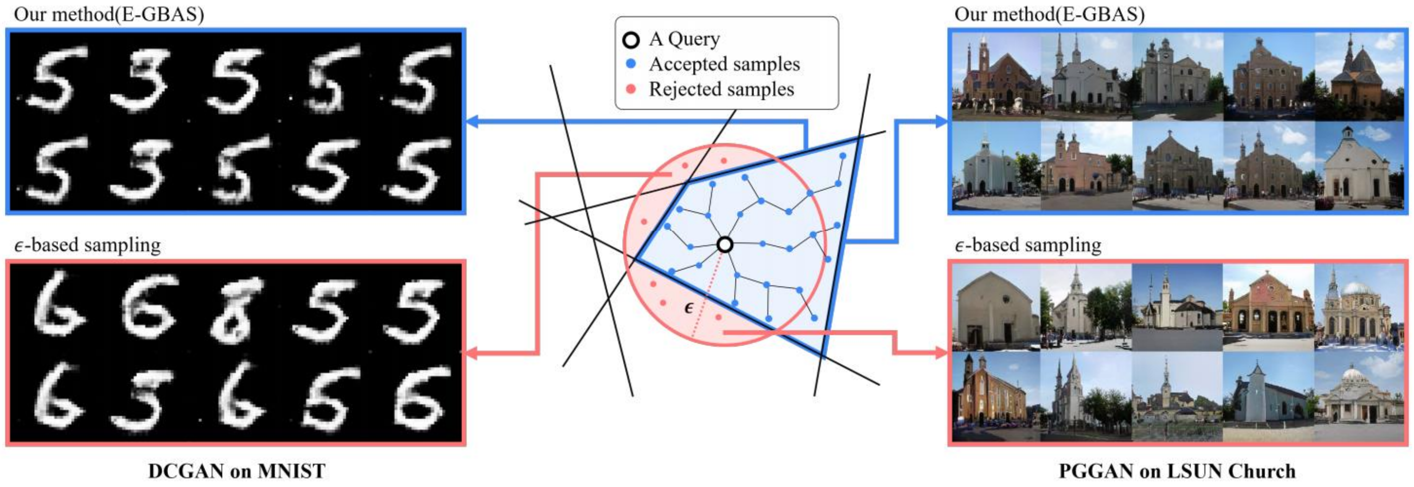
G. Jeon et. al., AAAI, 2020

A Generative Neural Network (GNN)





Generative boundaries

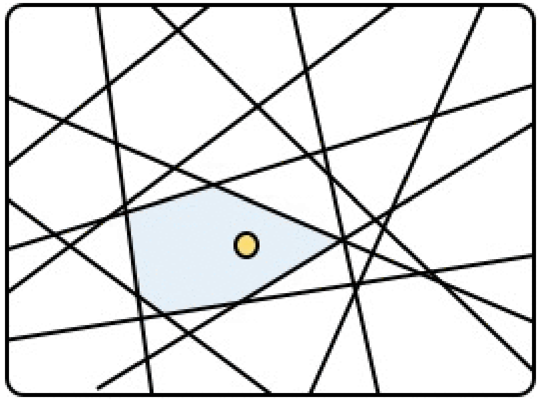


Explorative Generative Boundary Aware Sampling (E-GBAS)

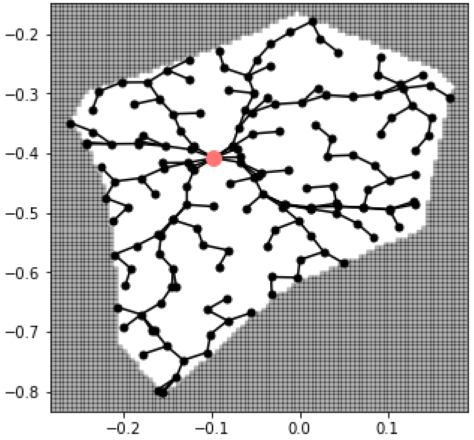
G. Jeon et. al., AAAI, 2020

Generative Boundary constrained Rapidly-exploring Random Tree (RRT)

- ☐ Given generative boundary as constraints,
  RRT is gives solution to search over the generative region.
- ☐ This explorative sampling always guarantee acceptance inside the region



Illustrative example



Example in nonconvex region

S. M. LaValle, “Rapidly-exploring random trees: A new tool for path planning”, 1998.

Explorative Generative Boundary Aware Sampling (E-GBAS)

G. Jeon et. al., AAAI, 2020

Explorative Generative Boundary Aware Sampling

- Accepted Cluster 1
- Accepted Cluster 2
- Accepted Cluster 3
- Rejected Sample

Explorative Generative Boundary Aware Sampling (E-GBAS)

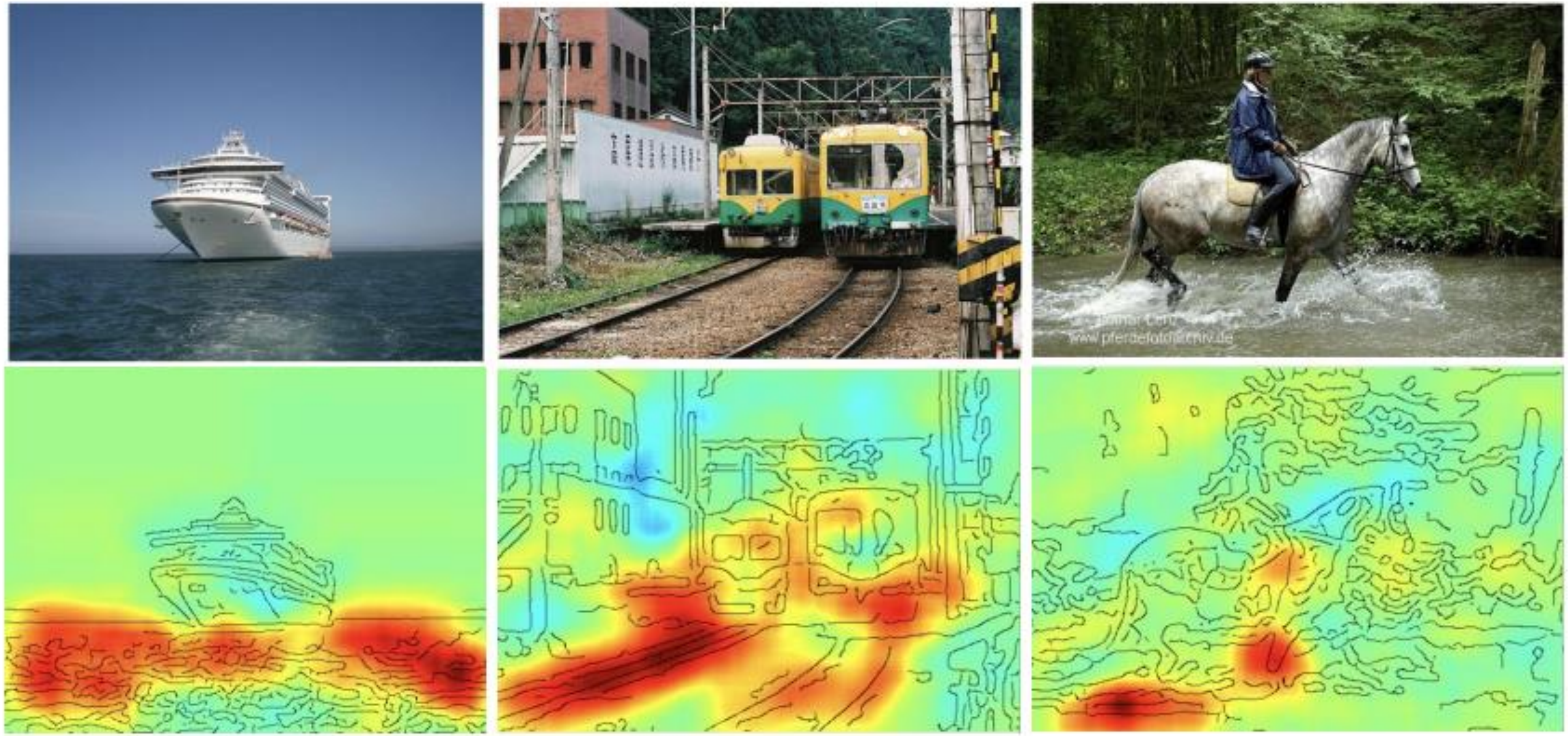
G. Jeon et. al., AAAI, 2020

Leading method (Fisher-Vector / SVM Model) of PASCAL VOC challenge

Unmasking Clever Hans Predictors

W. Samek, Unmasking Clever Hans Predictors, Nature Communications, 2019

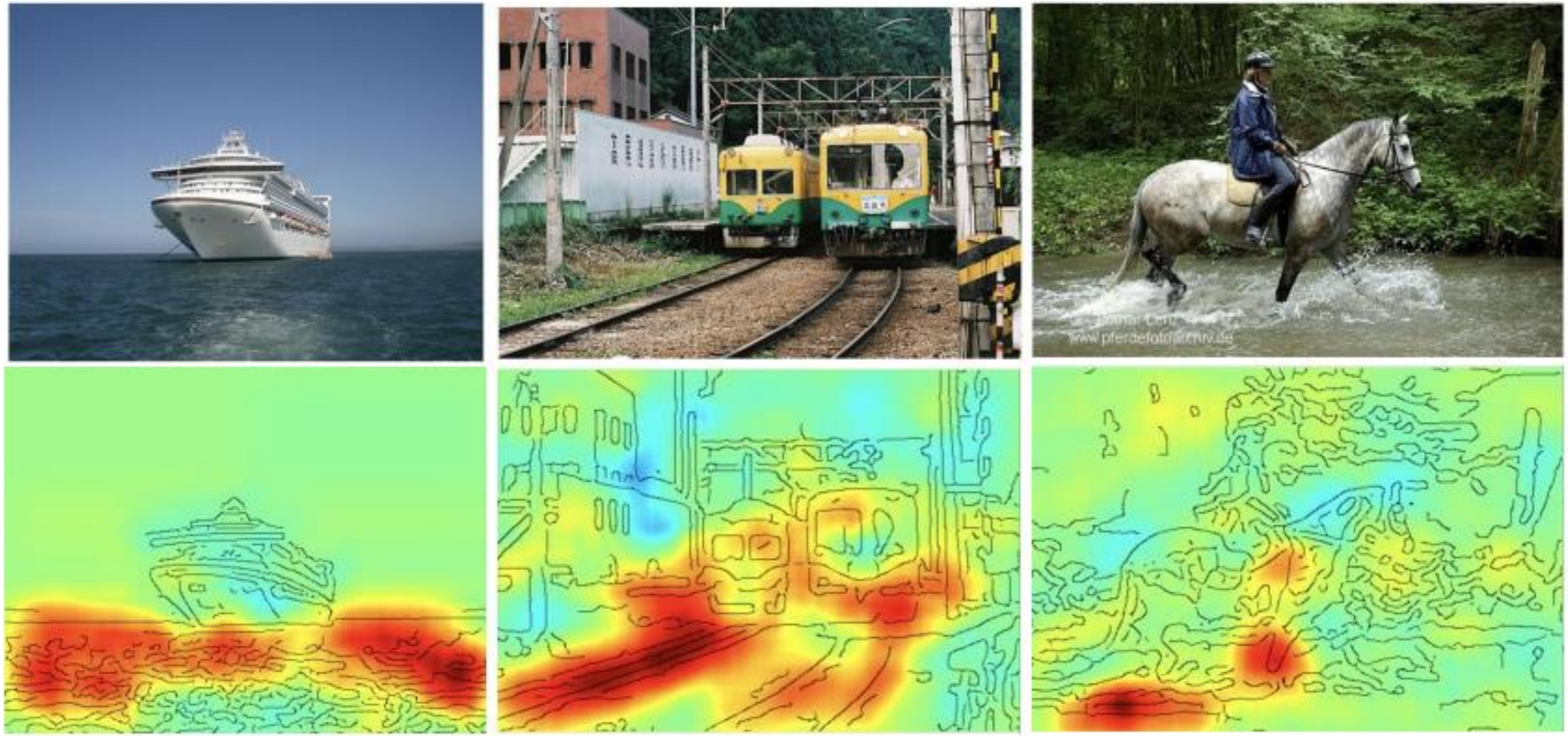
Leading method (Fisher-Vector / SVM Model) of PASCAL VOC challenge



Unmasking Clever Hans Predictors

W. Samek, Unmasking Clever Hans Predictors, Nature Communications, 2019

Leading method (Fisher-Vector / SVM Model) of PASCAL VOC challenge



Unmasking Clever Hans Predictors

W. Samek, Unmasking Clever Hans Predictors, Nature Communications, 2019

This is a ***mallard*** because this is a brown and white bird with a green head and a yellow beak.

Explainable AI can improve the accuracy of AI system

T. Darrell, Recent progress towards XAI at UC Berkeley, 2019

This is a ***mallard*** because this is a brown and white bird with a green head and a yellow bill.

Explainable AI can improve the accuracy of AI system

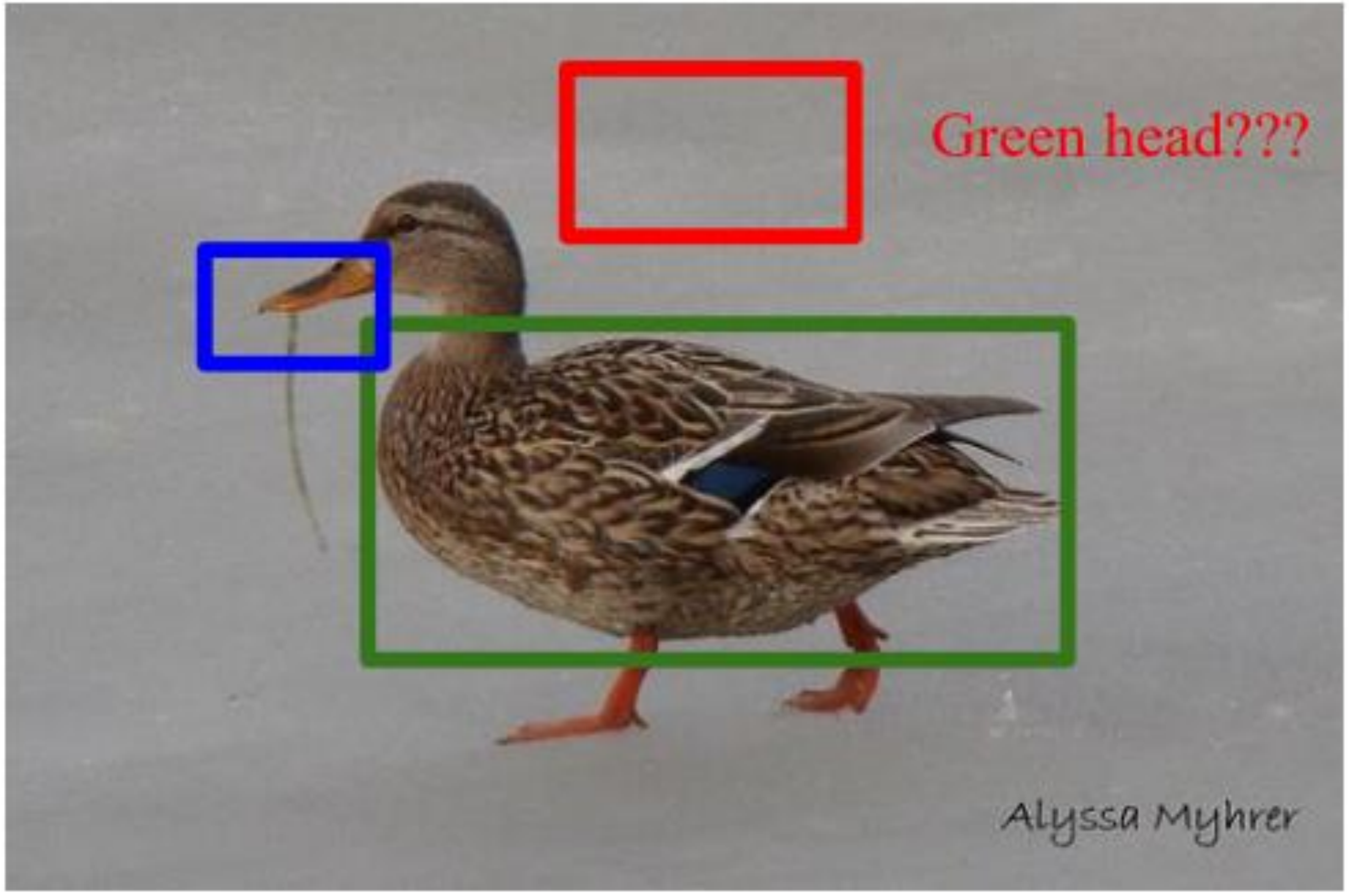
T. Darrell, Recent progress towards XAI at UC Berkeley, 2019

~~This is a ***mallard*** because this is a brown and white bird with a green head and a yellow bill.~~

This is a ***mallard*** because this bird has a brown head, orange feet, and a flat bill.

Explainable AI can improve the accuracy of AI system

T. Darrell, Recent progress towards XAI at UC Berkeley, 2019



This is a ***mallard*** because this is a ***brown and white bird*** with a ***green head*** and a ***yellow bill***.

Explainable AI can improve the accuracy of AI system

T. Darrell, Recent progress towards XAI at UC Berkeley, 2019

- Interpretable and explainable AI methods are **necessary for the coexistence of human and AI**.

- Recent advances in XAI can **analyze internal nodes of deep neural networks**.

- Some XAI methods help to **improve the performance of AI systems**.

Conclusions

Conditions for Trustworthy AI:
Explainable Artificial Intelligence

Jaesik Choi

Explainable Artificial Intelligence Center
Graduate School of Artificial Intelligence
KAIST