PyCon Korea 2019

딥러닝 안에서 일어나는 과정을 해석하는 설명가능인공지능 기술

2019년 8월 16일

UNIST 설명가능인공지능 연구센터
허성만, 조소희, 최재식

진행

최재식

UNIST ECE 부교수
UNIST XAIC 센터장

조소희

UNIST XAIC 연구원

허성만

UNIST XAIC 연구원

보조 진행자

차성국

UNIST Sailab

이희찬

UNIST Sailab

목차

Recent Advances in Explainable Artificial Intelligence

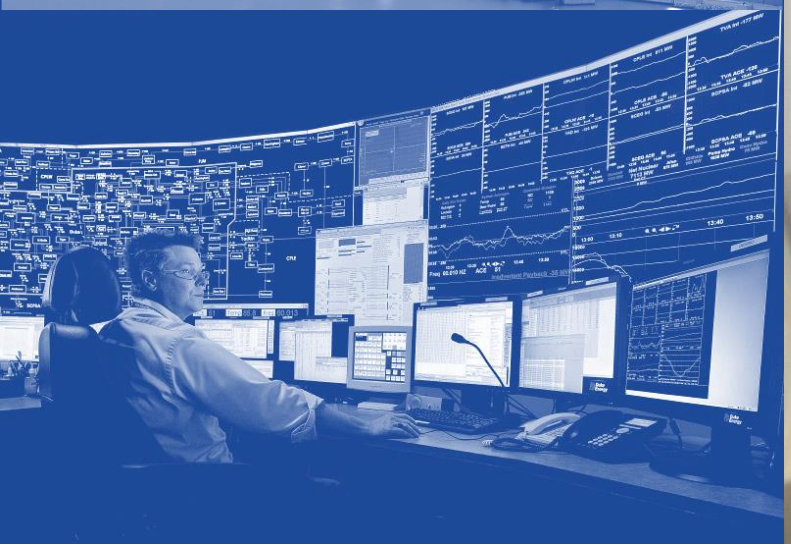
Jaesik Choi*

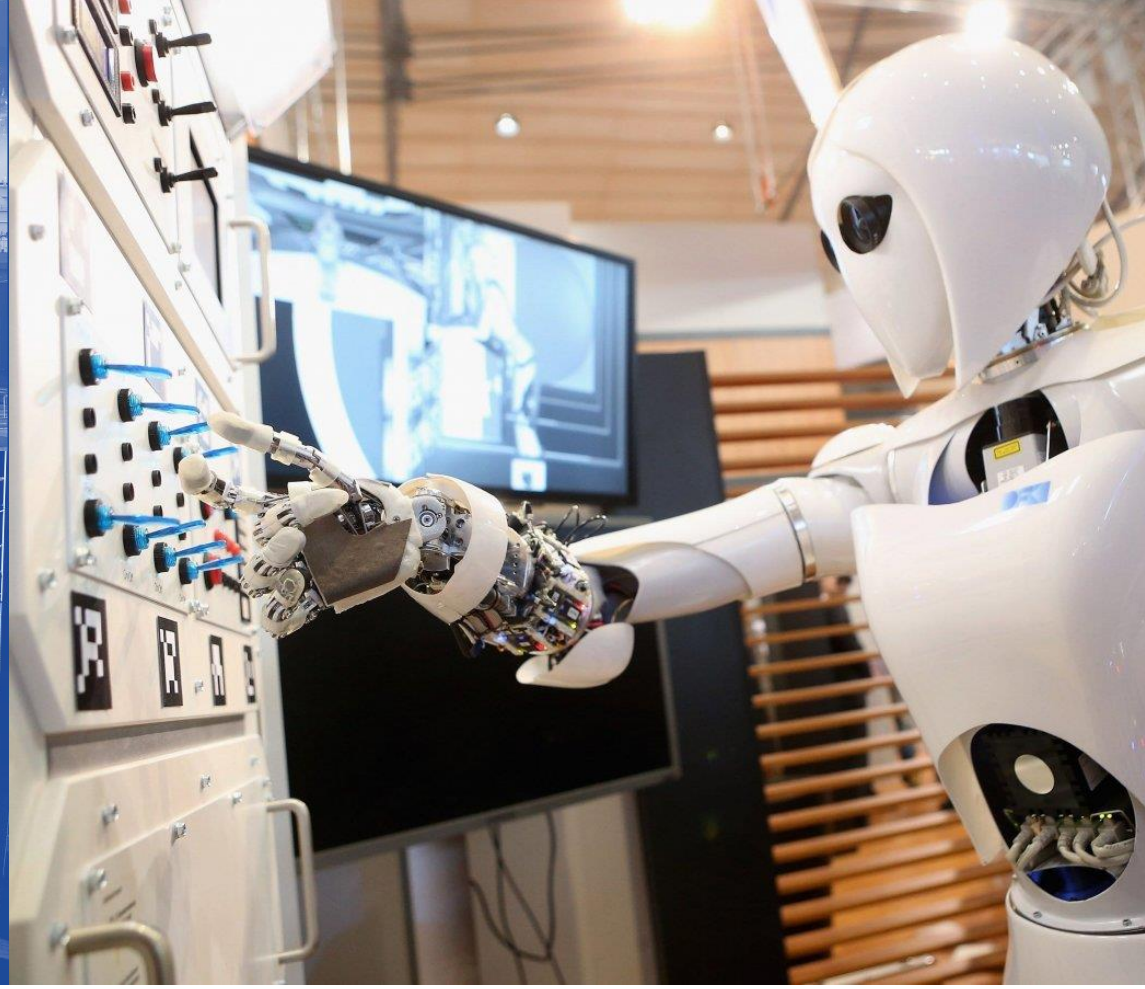
Explainable Artificial Intelligence Center (XAIC)

Statistical Artificial Intelligence Lab (SAIL)

UNIST

* Joint work with Anh Tong, Yunseong Hwang, Yeeun Chun, Sohee Cho, Janghoon Ju, Sol-A Kim, Kyowoon Lee, Seong-Whan Lee





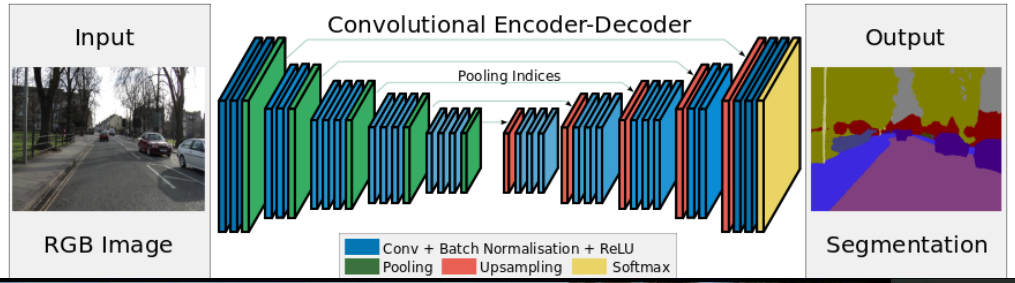
인공지능의 미래?

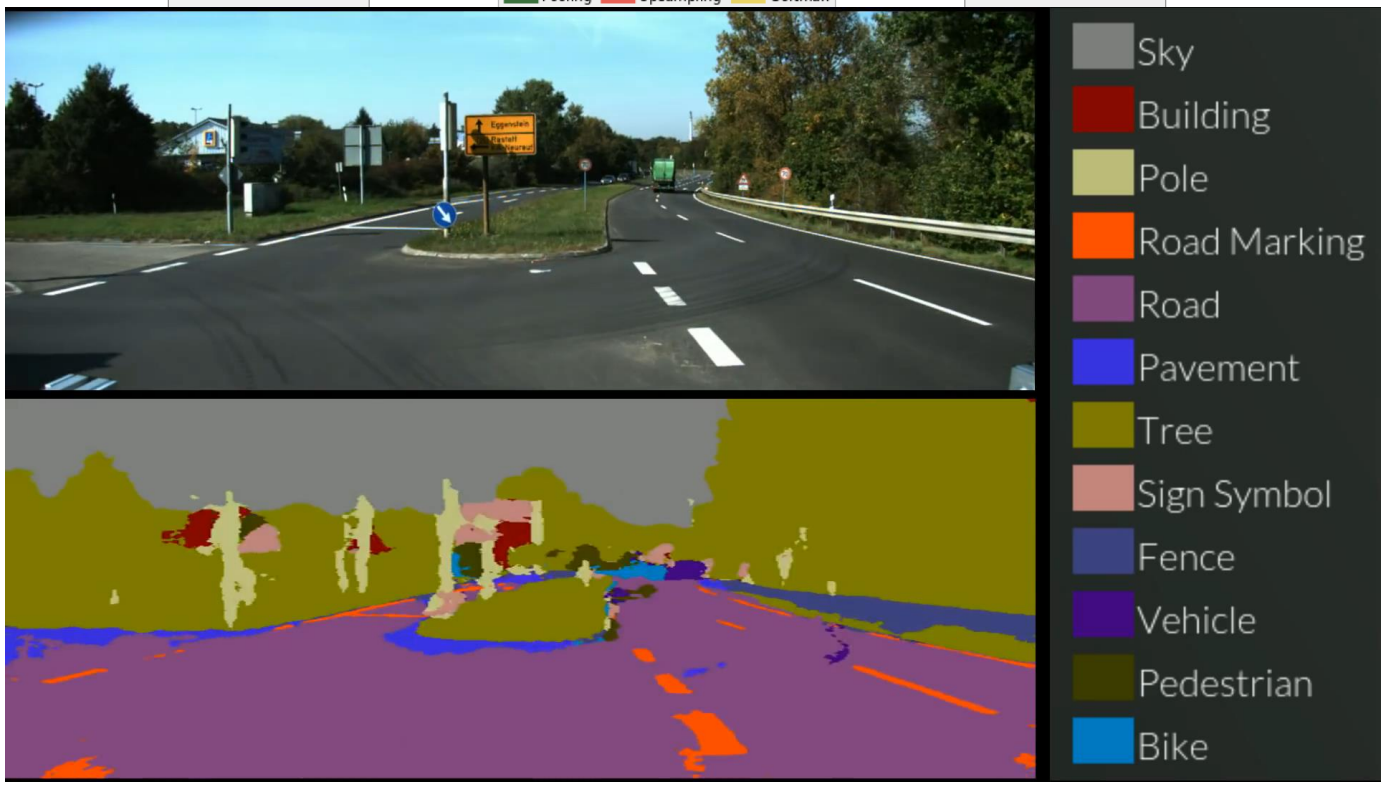
https://www.youtube.com/watch?v=7a6GrKqOxeU

DARPA Grand Challenge 2005

https://www.youtube.com/watch?v=7a6GrKqOxeU

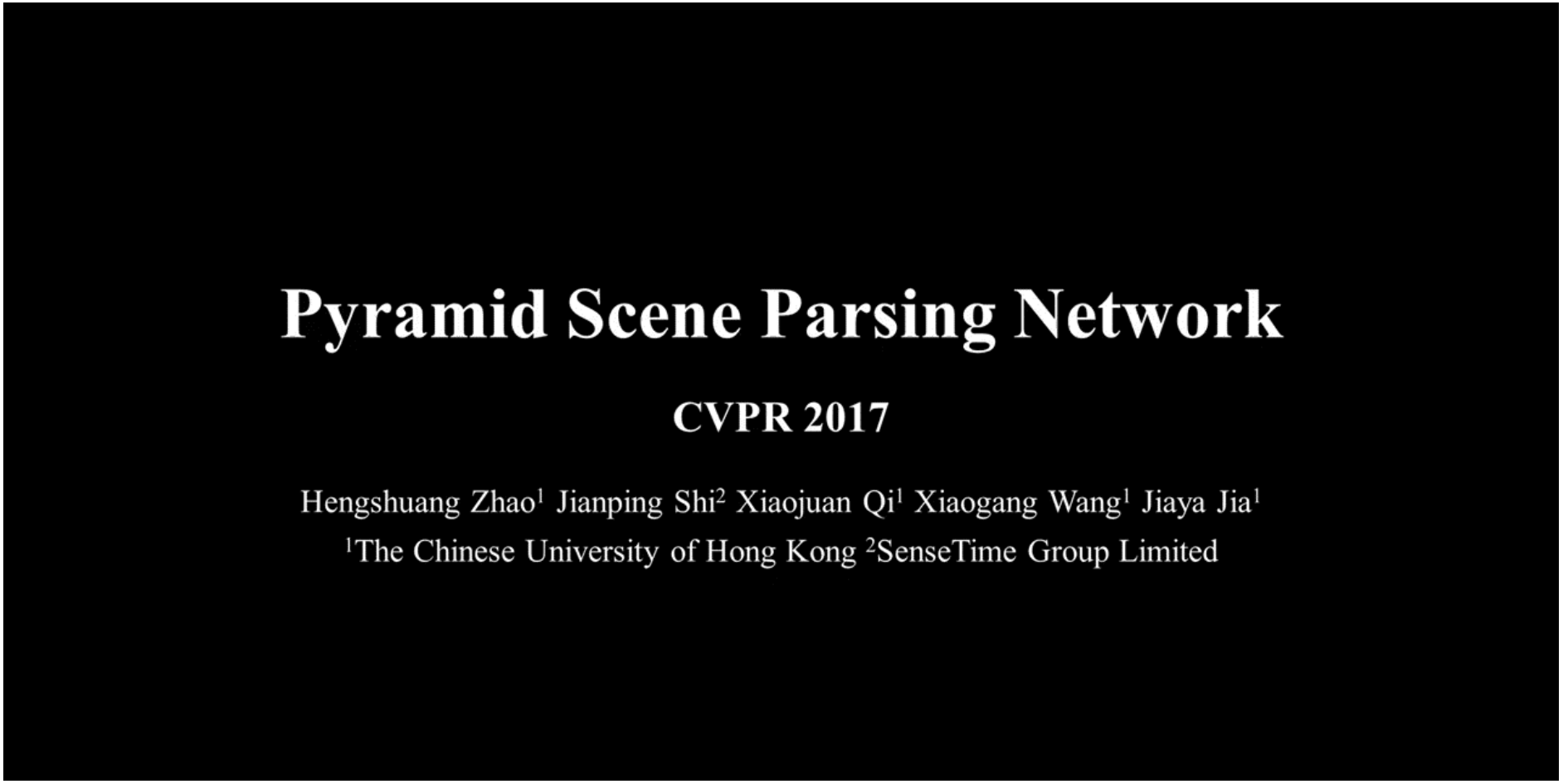
Say Hello to Waymo 2016





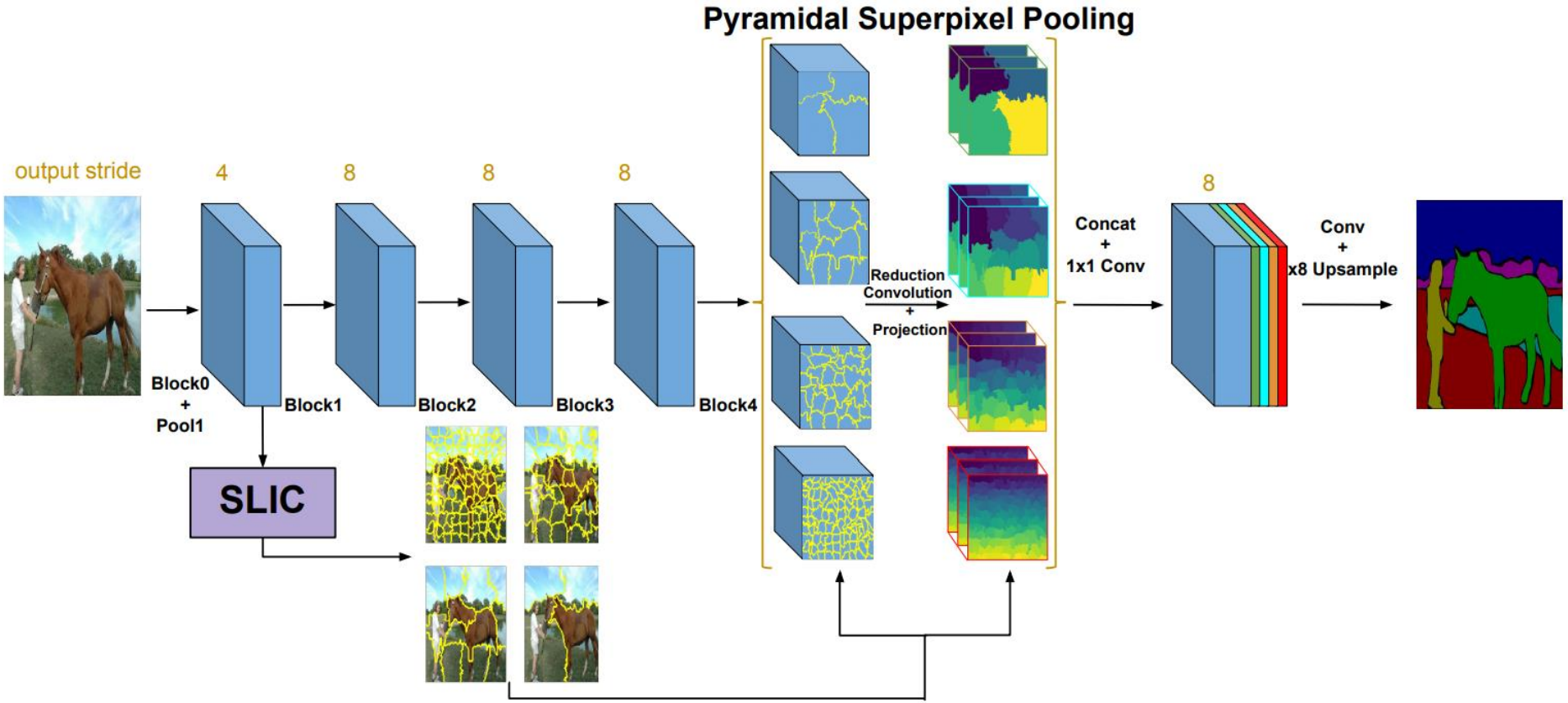
https://www.youtube.com/watch?v=CrnLINlbfFA

Semantic Segmentation by SegNet 2015



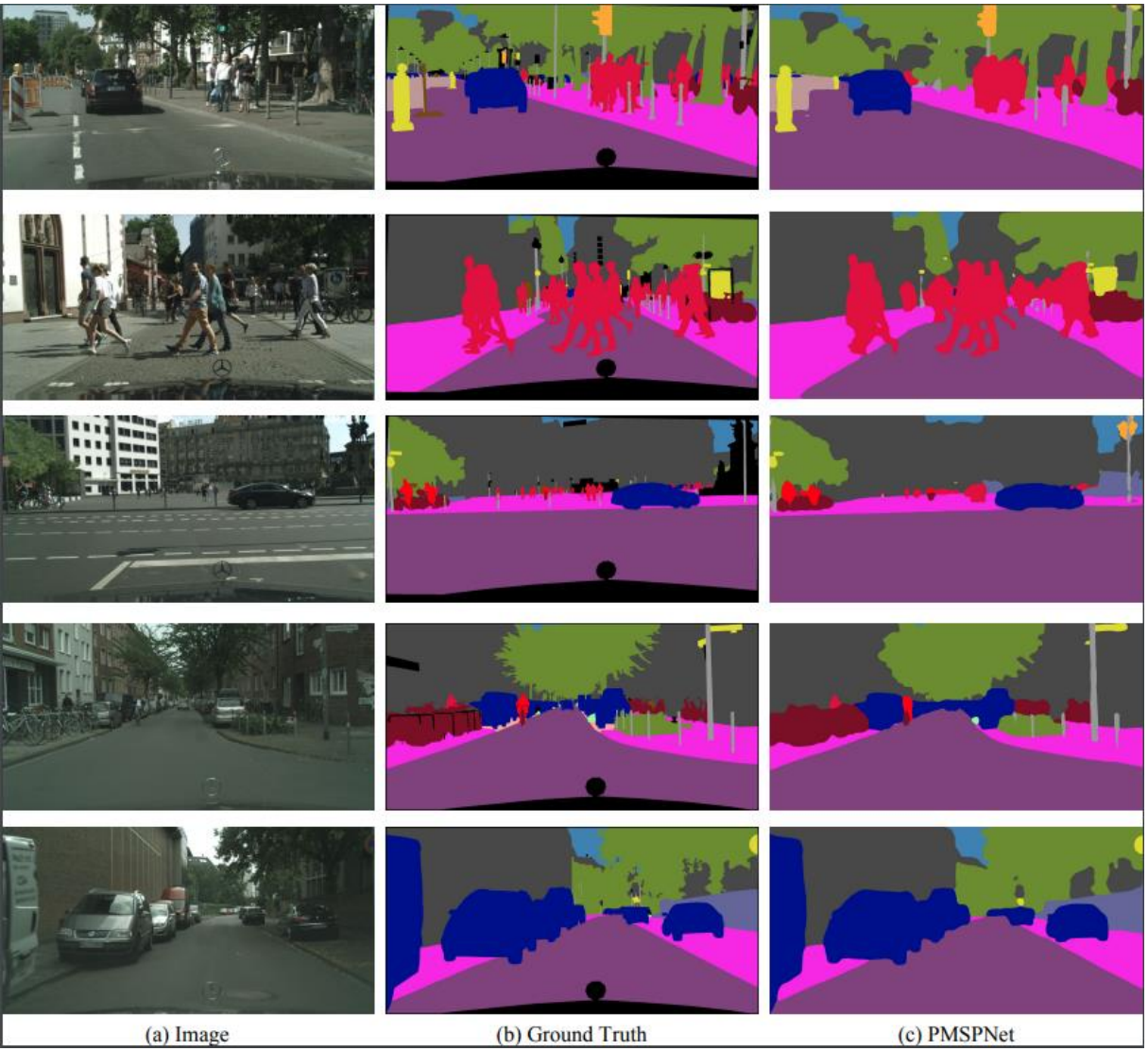
https://www.youtube.com/watch?v=gdAVqJn_J2M

Semantic Segmentation by Pyramid Scene Parsing Network 2017



Ali Tousi and Jaesik Choi, Pyramidal Multiscale Superpixel Pooling Network for Semantic Segmentation, soon will be available at arxiv.

Pyramid Multiscale Superpixel Pooling Network 2019



Method	mIoU
DeepLab-v2 [3]	70.4
LC [14]	71.1
Adelaide_context [16]	71.6
FRRN [22]	71.8
RefineNet[15]	73.6
FoveaNet [13]	74.1
PEARL [9]	75.4
DUC_HDC [29]	77.6
SAC_multiple [37]	78.1
PSPNet†[38]	78.4
ResNet-38 [31]	78.5
SegModel [26]	78.5
Multitask Learning [10]	78.5
PSANet [40]†	78.6
PSANet [40] ‡	**80.1**
PMSPNet†(ours)	78.6
PMSPNet‡(ours)	**80.4**

† Trained with *fine_train* set only

‡ Trained with both *fine_train* and *fine_val* set

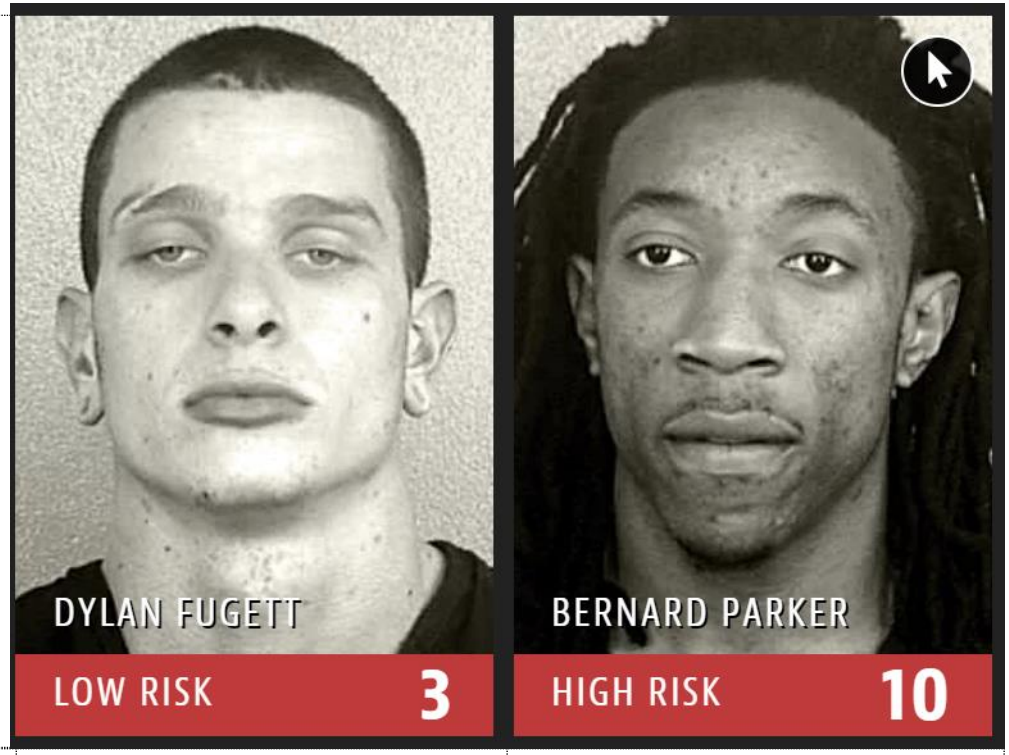
Pyramid Multiscale Superpixel Pooling Network 2019

우버 자율주행차 첫 보행자 사망사고

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

Prior Offense	1 attempted burglary	1 resisting arrest without violence
COMPAS' decision		
Subsequent Offenses	3 drug possessions	None

Do We Understand AI Systems Enough?

금융

유럽연합 General Data Protection Regulation 발효(2018년)
- 인공지능 알고리즘에 의해 자동으로 결정된 사안에 대해 회사의 설명을 강제

미국 Equal Credit Opportunity Act/Fair Housing Act
- 신용결정 및 주택 담보 대출 등 주요 금융 결정에 대해서 이유를 제시하도록 강제

현 인공지능 기반 신용 평가의 장단점

장점	단점
세밀한 분석을 통한 정교한 신용 점수 산정으로, **숨어있는 우수 고객을 선발** 할 수 있음	신용 거부의 **이유를 분명히 제시하지 못하여,** 인공지능 모델이 감독기관의 승인을 못 받을 수 있음

MIT Technology Review

Intelligent Machines

The Financial World Wants to Open AI's Black Boxes

THE WALL STREET JOURNAL.

Capital One Pursues 'Explainable AI' to Guard Against Bias in Models

시사점

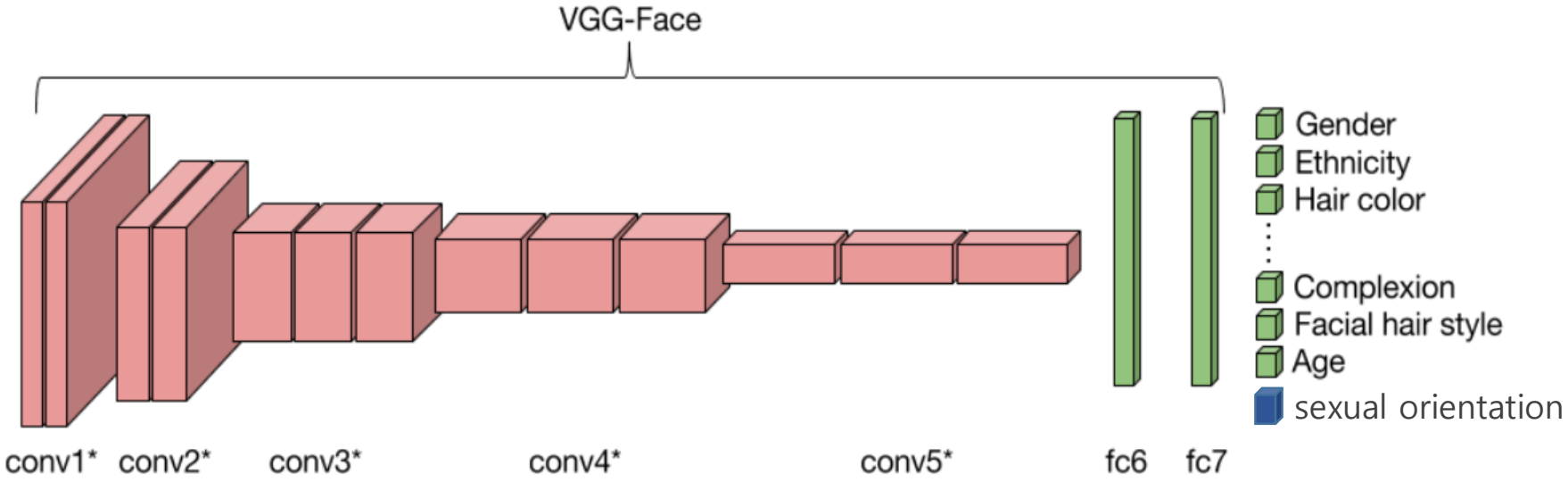
인공지능 기술을 주요 의사결정에 적용하기 위해서 법적으로 이유 제시 기능이 필요함

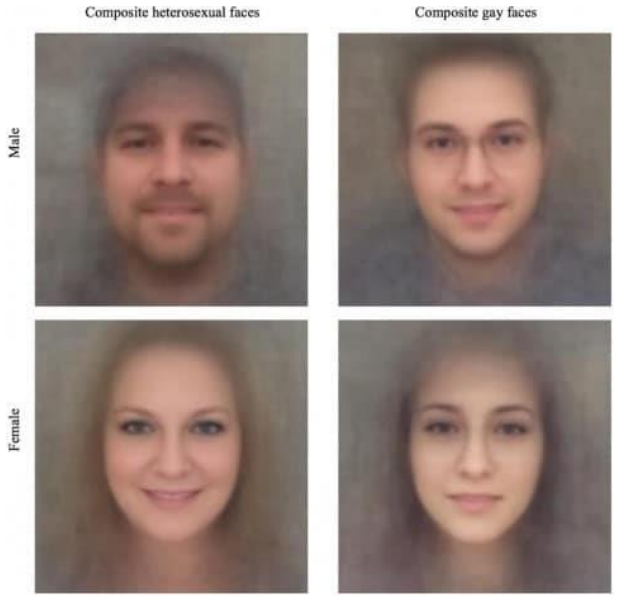
https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing

EU의 일반정보보호규정(General Data Protection Regulation)

항목	내용
잊혀질 권리 **(right to be forgotten)**	제17조 – 정보 주체가 본인의 개인정보 처리를 더 이상 원치 않거나 개인정보를 보유할 법적 근거가 없으면 **해당 정보 삭제**
자동화된 의사결정 제한	제22조 – **자동화된 처리**(프로파일링 포함)에만 근거한 결정의 대상이 **되지 않을 권리**
설명을 요구할 권리 **(right to explanation)**	제13-14조 – 알고리즘에 의해 행해진 결정에 대해 질문하고, **결정에 관여한 논리에 대해 의미있는 설명을 요구할 권리**
EU 집행력	규정 위반시 해당 기업의 **전세계 매출의 최대 4%**까지 벌금 부과
발효	**2018년 5월 28일**

사례 I - 사람의 성적 지향을 판별하는 딥러닝?





기존 동성애자/양성애자 판별 정확도: 55 ~ 65%

VS

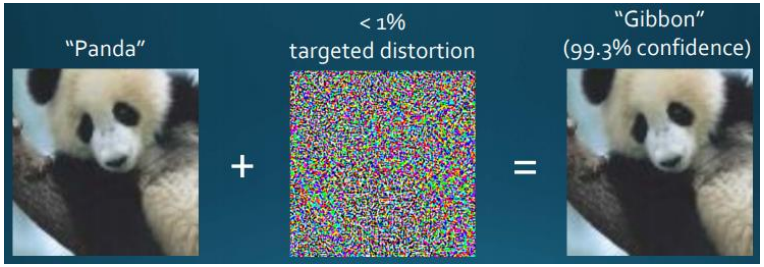
VGG-Face의 정확도: 83 ~ 91%

문제: VGG-Face는 결정의 이유를 설명할 수 없음!

M. Kosinski 교수 연구팀 (Stanford), 2017년

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

Statistically impressive, but individually unreliable

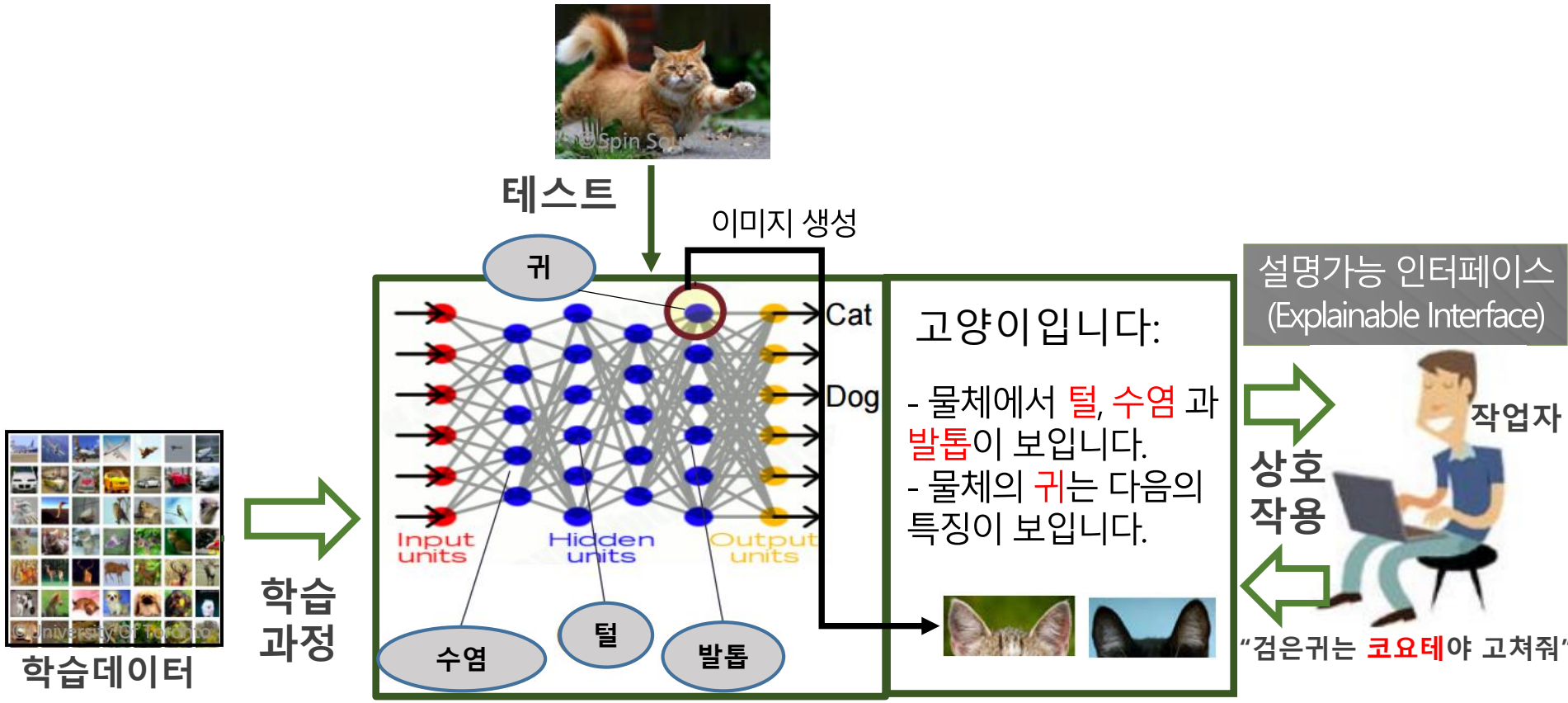


Inherent flaws can be exploited

Skewed training data creates Maladaptation

A DARPA Perspective on AI – Three Waves of AI

해결방법 – 설명을 제공하는 인공지능 시스템



<설명 가능한 딥러닝 모델의 예>

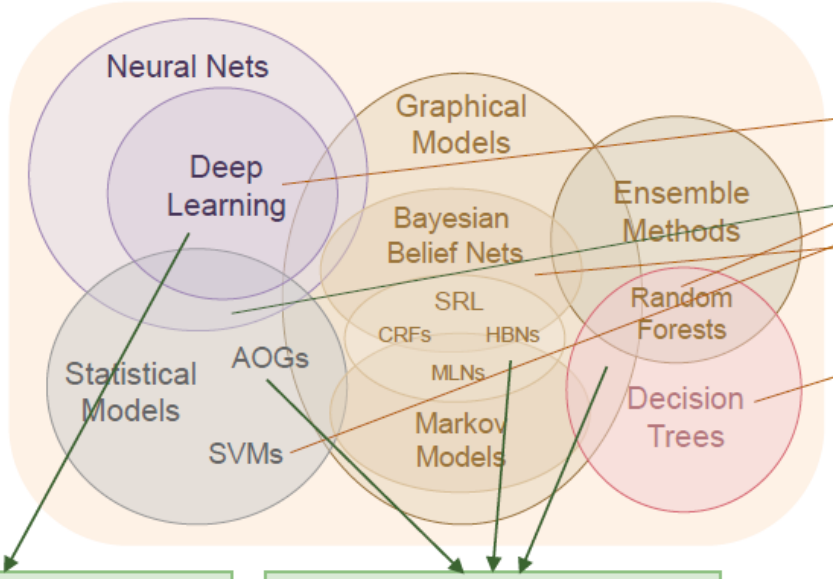
해결방법 - DARPA 설명가능 인공지능

Explainable AI – Performance vs. Explainability

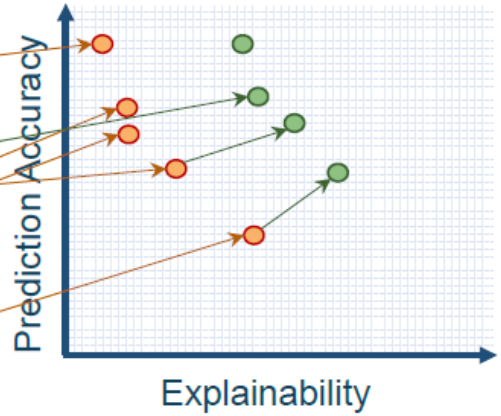
New Approach

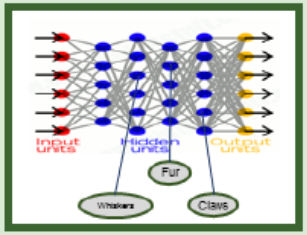
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



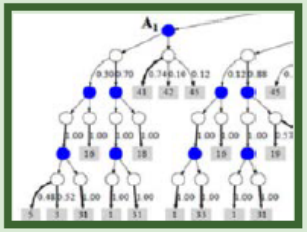
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



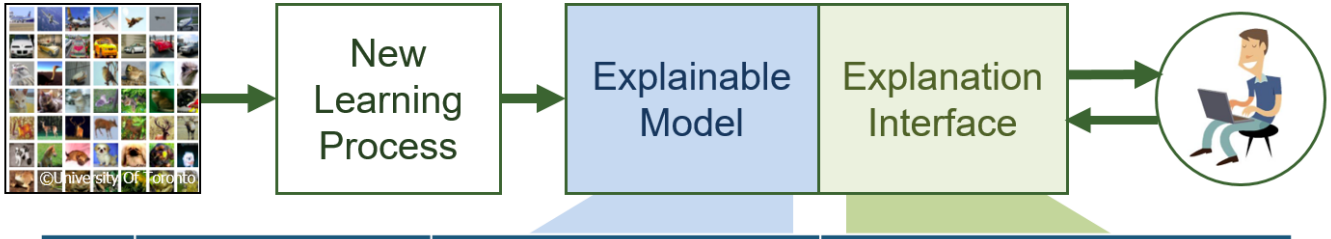
Interpretable Models

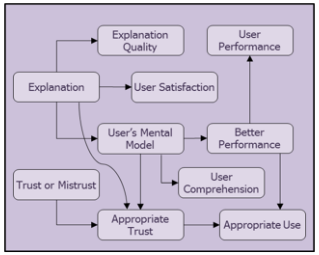
Techniques to learn more structured, interpretable, causal models

D. Gunning (DARPA), 2016년

해결방법 - DARPA 설명가능 인공지능

XAI Performers & Technical Approaches





IHMC Psychological Models of Explanation

CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

UNIST Explainable AI Research Center

Partners

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Technische Universität Berlin
UC DAVIS UNIVERSITY OF CALIFORNIA
JOHNS HOPKINS UNIVERSITY
IBM WATSON
MICHIGAN STATE UNIVERSITY
Penn
MAX PLANCK GESELLSCHAFT
Microsoft Research
MIT
UCLA
SRI International
Carnegie Mellon University

AI algorithms and applications that provide explanations

http://www.openXAI.org/

SW release and Technology promotion

Technology & DB Cooperation

Explainable AI Research Center

AITRICS
KAIST 1971
KOREA UNIVERSITY 1905
UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY 2009
YONSEI UNIVERSITY 1885
VERITAS LUX MEA

Technology Sharing
Competition

kaggle
UNIST 제6회 빅데이터 분석 경진대회
Rotman-UNIST Trading Competition 2017
International Finance Trading / Medical Big Data Competition

Obtain medical data & Evaluate usability

Medical application

YONSEI UNIVERSITY 세브란스 SEVERANCE
UNIST 빅데이터연구센터

Obtain financial data & Evaluate usability

Financial application

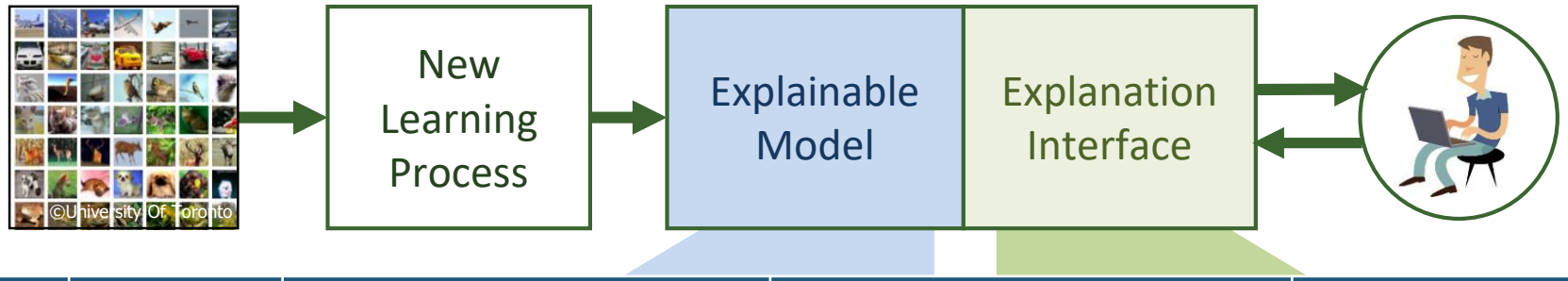
자본시장 IT 파트너
koscom
UNIST 국제에너지트레이딩연구센터

Demand for technology
Technology transfer & commercialization

Enterprises with intention of technology transfer

NAVER
AI ZEN AI for Better Finance
posco
BNK 경남은행

XAI Performers & Technical Approaches



Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning	Textual Explanation	Visual Dialog
Task 3	UNIST	Nonparametric Bayesian	Narrative Explanation	Finance
	Korea University	Deep Learning (attribution methods)	Visual Explanation	Healthcare
	KAIST/ UNIST	Explainable RL	Interactive Visualization	Game

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning

Explaining by Combining Explainable Models

Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Technical Approaches

Explaining Deep Neural Networks

- Input Attribution Method (TU Berlin)
- XCAD: Explainable Computer Aided Diagnosis (KAIST)
- Dissecting Deep Neural Networks (MIT)
- Generating Examples inside of Deep Neural Networks (UNIST)

Explaining Deep Reinforcement Learning

Explaining by Combining Explainable Models

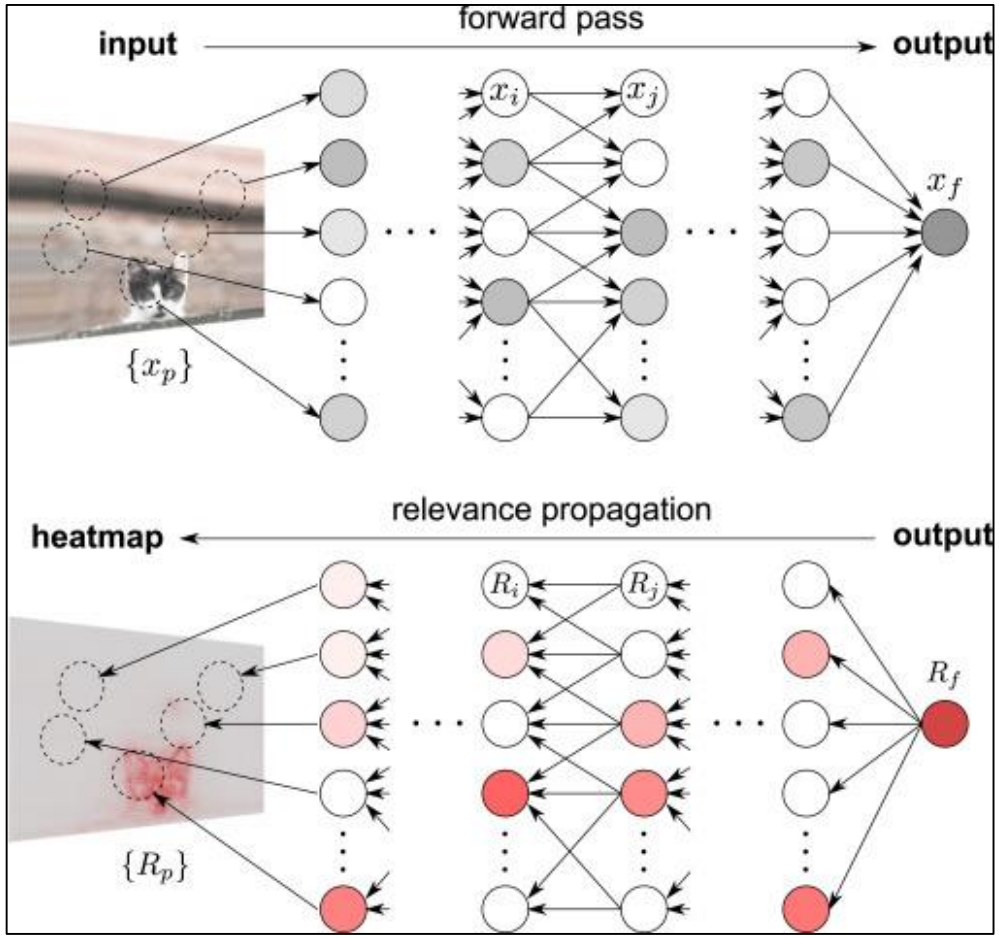
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Input Attribution Methods

Layer-wise Relevance Propagation(LRP)

딥러닝 모델의 예측 결과에 대해 시각적 해석을 돕는 공개 SW를 Heatmapping.org 에 공개

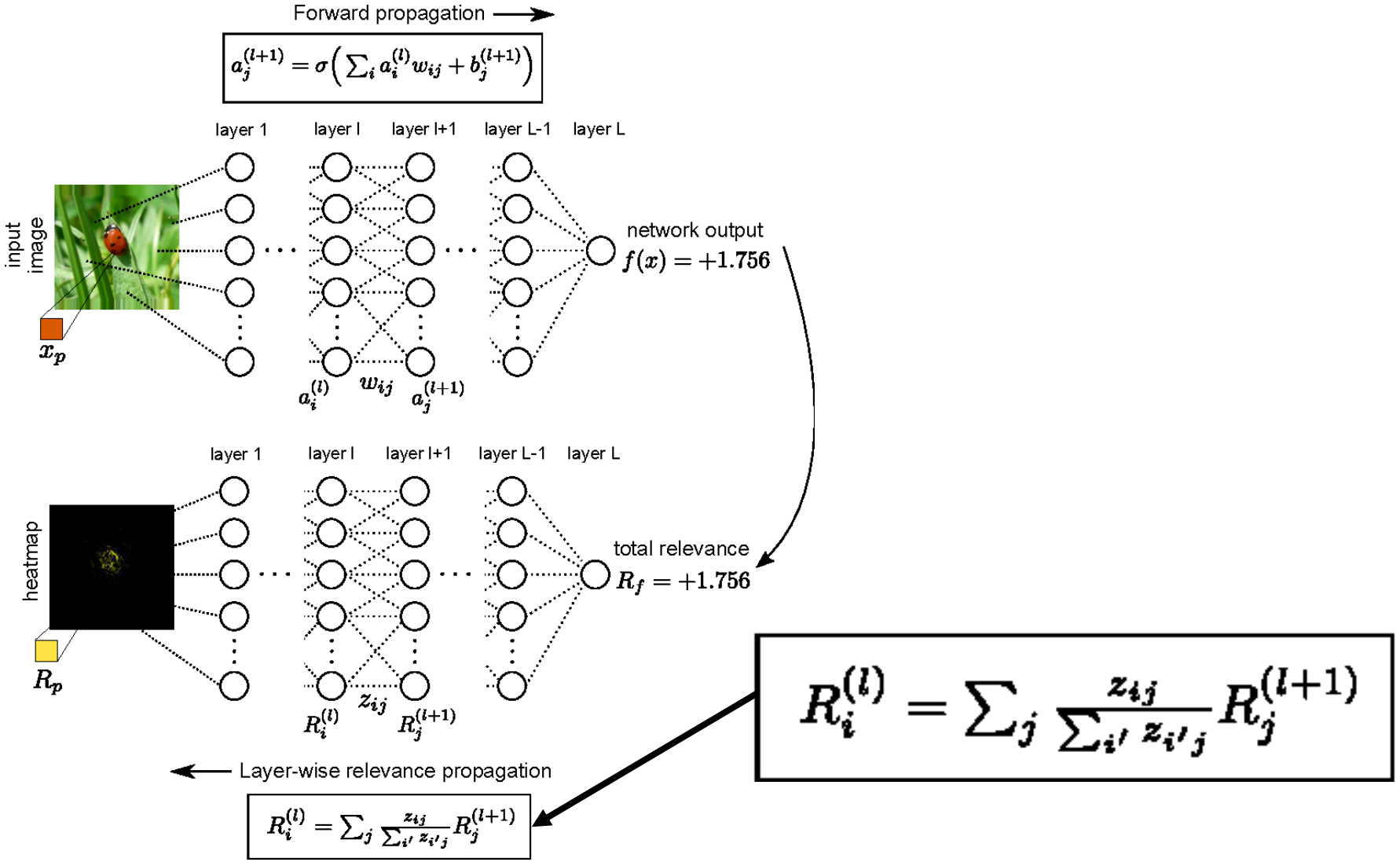


딥러닝 예측 결과 이유 설명을 위한 LRP 알고리즘 [CVPR, 2015]

K.-R. Muller 교수 연구팀 (고려대학교/Technical Univ. of Berlin), 2015년

Input Attribution Methods

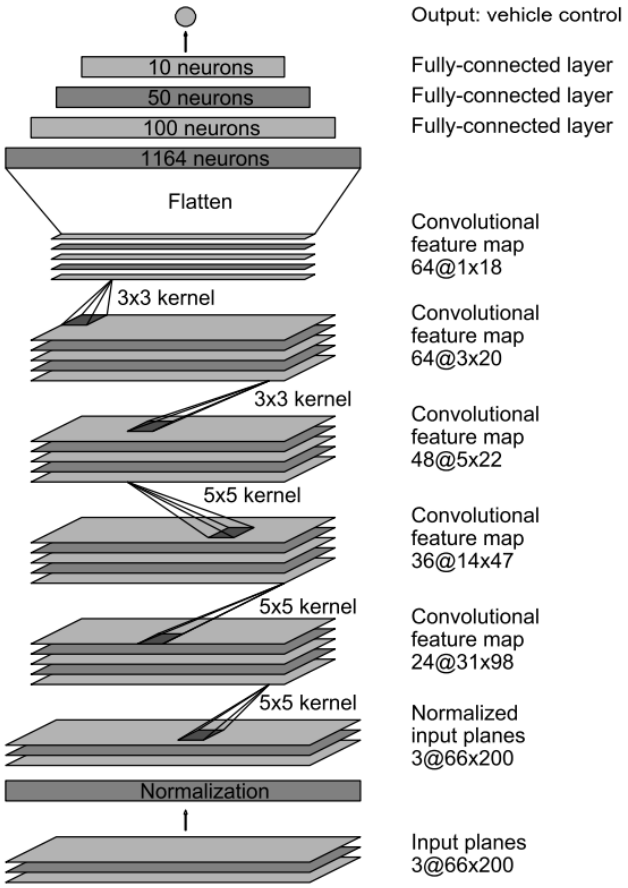
Layer-wise Relevance Propagation (LRP)



K.-R. Muller's Group (Korea University/Technical Univ. of Berlin), 2015

Input Attribution Methods

NVIDIA의 PilotNet(자율주행 딥러닝)의 결정을 설명하는 딥러닝



PilotNet 구조도

자율주행의 이유를 설명해주는 딥러닝 기술

NVIDIA/Google 연구팀, 2017년

Input Attribution Methods

Saliency Maps
Simonyan et al. 2015

Integrated Gradients
Sundararajan et al. 2017

DeepLIFT
Shrikumar et al. 2017

LIME
Ribeiro et al. 2016

Gradient * Input
Shrikumar et al. 2016

Layer-wise Relevance Propagation (LRP)
Bach et al. 2015

Guided Backpropagation
Springenberg et al. 2014

Grad-CAM
Selvaraju et al. 2016

Simple occlusion
Zeiler et al. 2014

Meaningful Perturbation
Fong et al. 2017

Prediction Difference Analysis
Zintgraf et al. 2017

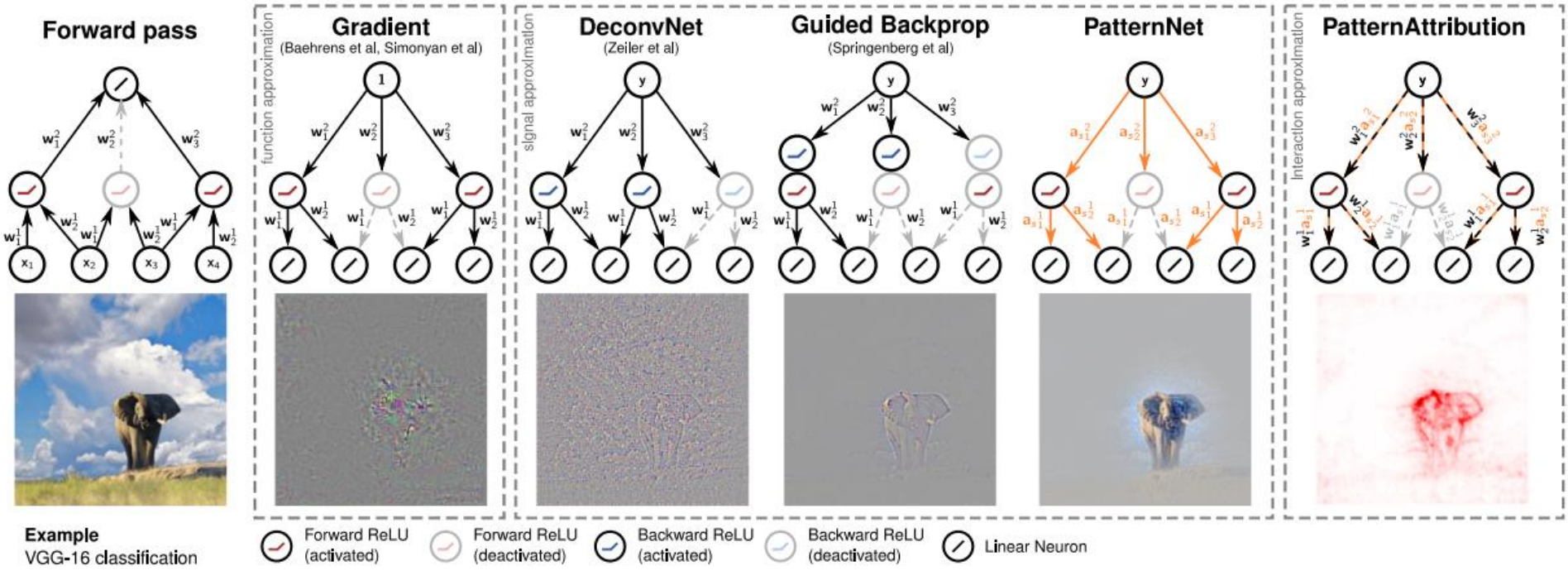
KernelSHAP/DeepSHAP
Lundberg et al., 2017

Slides courtesy of [Marco Ancona, et. al., Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation, ICML 2019]

PatternNet – Learning to Explain Neural Net – 1/2

Klaus-Robert Müller (TU Berlin/Korea University)

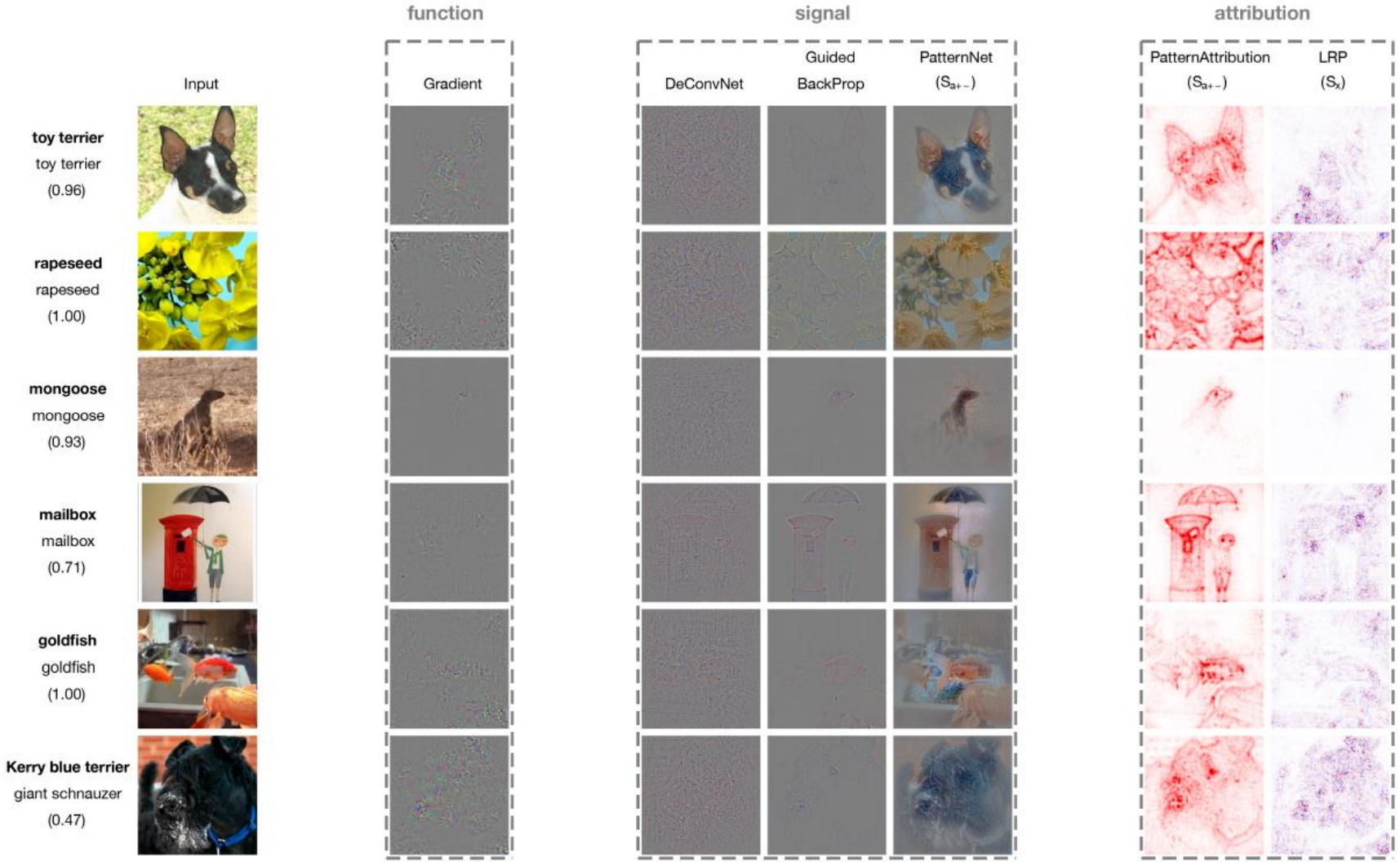
Learning (Optimizing) Attributions to Reduce Signal Noise



Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018

PatternNet – Learning to Explain Neural Net – 2/2

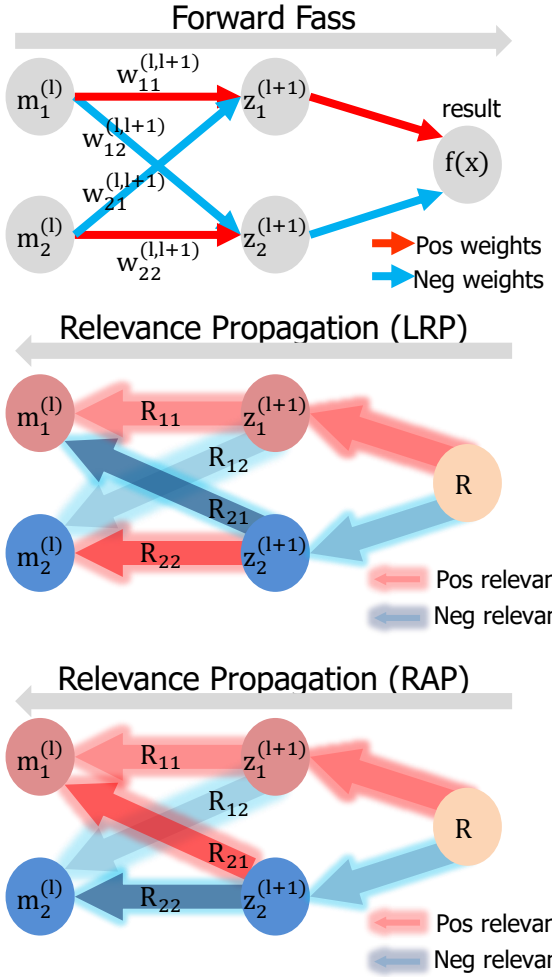
Klaus-Robert Müller (TU Berlin/Korea University)



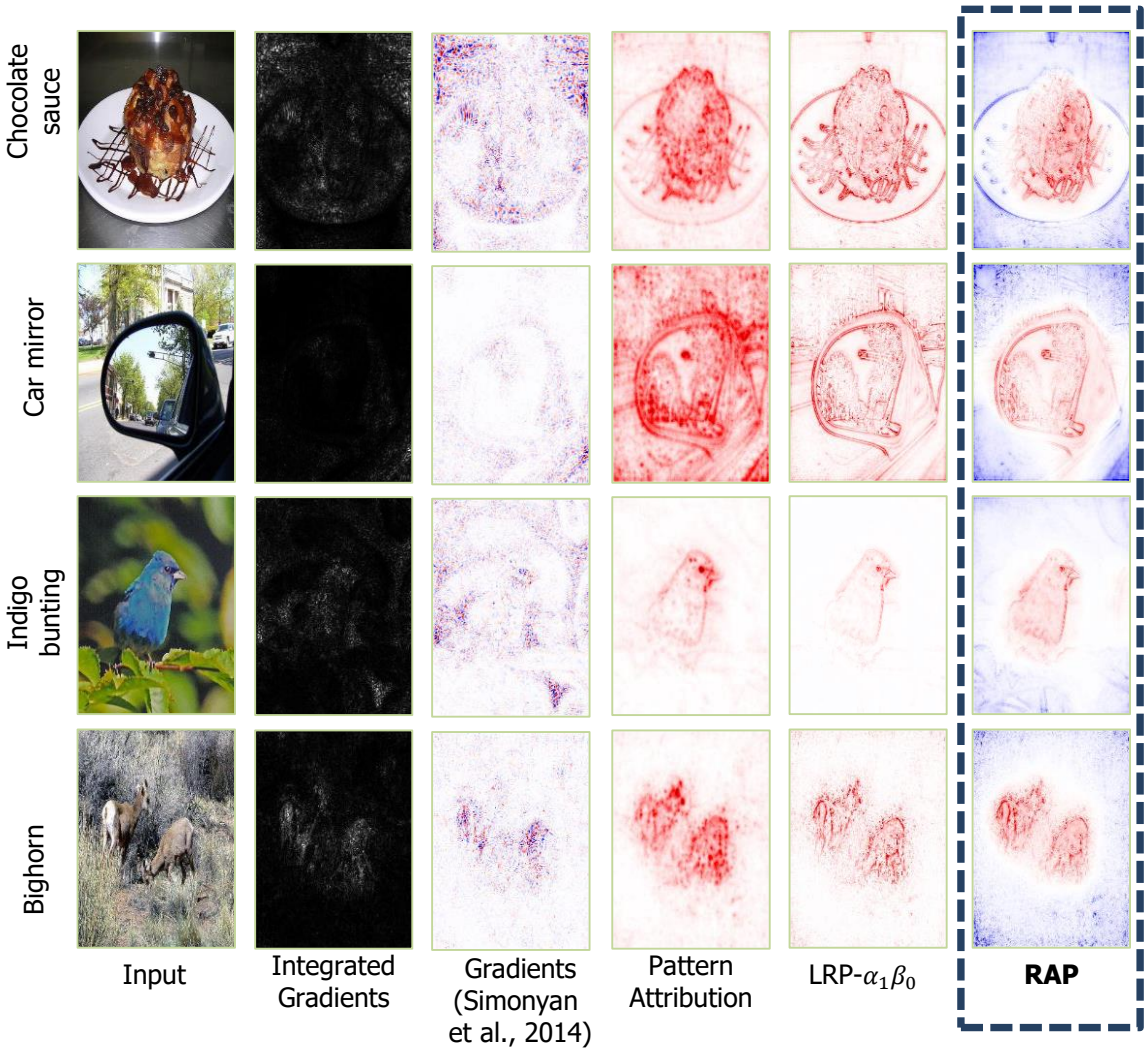
Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018

RAP - Relative Attribution Propagation

Seong-Whan Lee (Korea University)
Jaesik Choi (UNIST)



Propagating flipped attributions of negative relevance

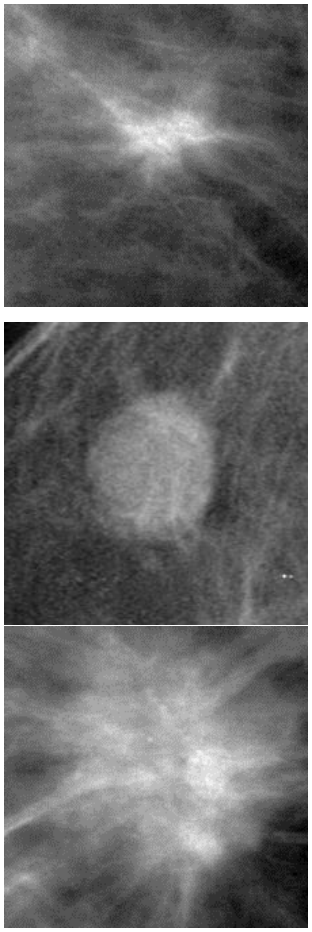


Woo-Jeong Nam, Jaesik Choi, and Seong-Whan Lee, "Relative Attribution Propagation", arXiv 1904.00605, 2019

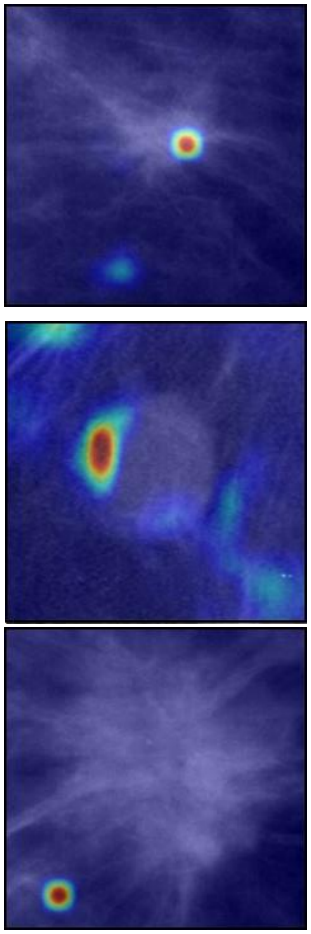
XCAD: Explainable Computer Aided Diagnosis – 1/2

Yong Man Ro (KAIST)

Test image



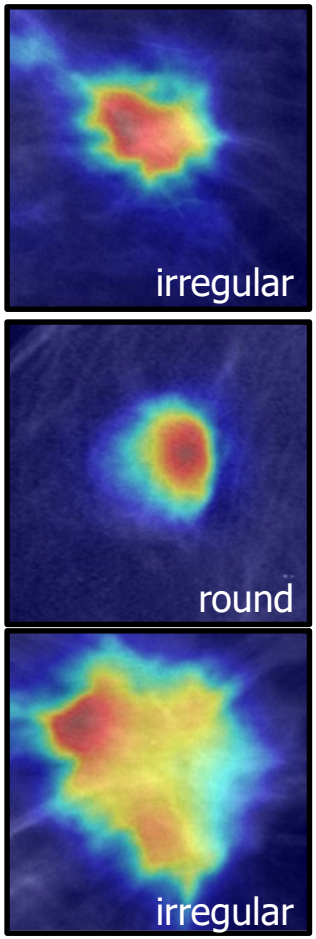
grad-CAM



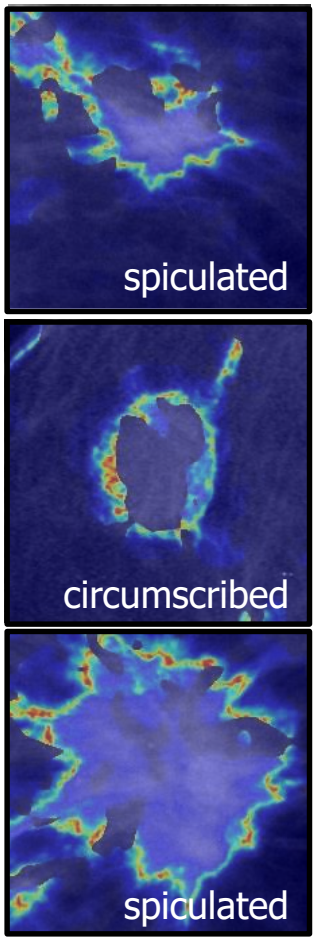
Difficult for medical doctors

Explanations of XCAD guided by BI-RADS*

Shape



Margin



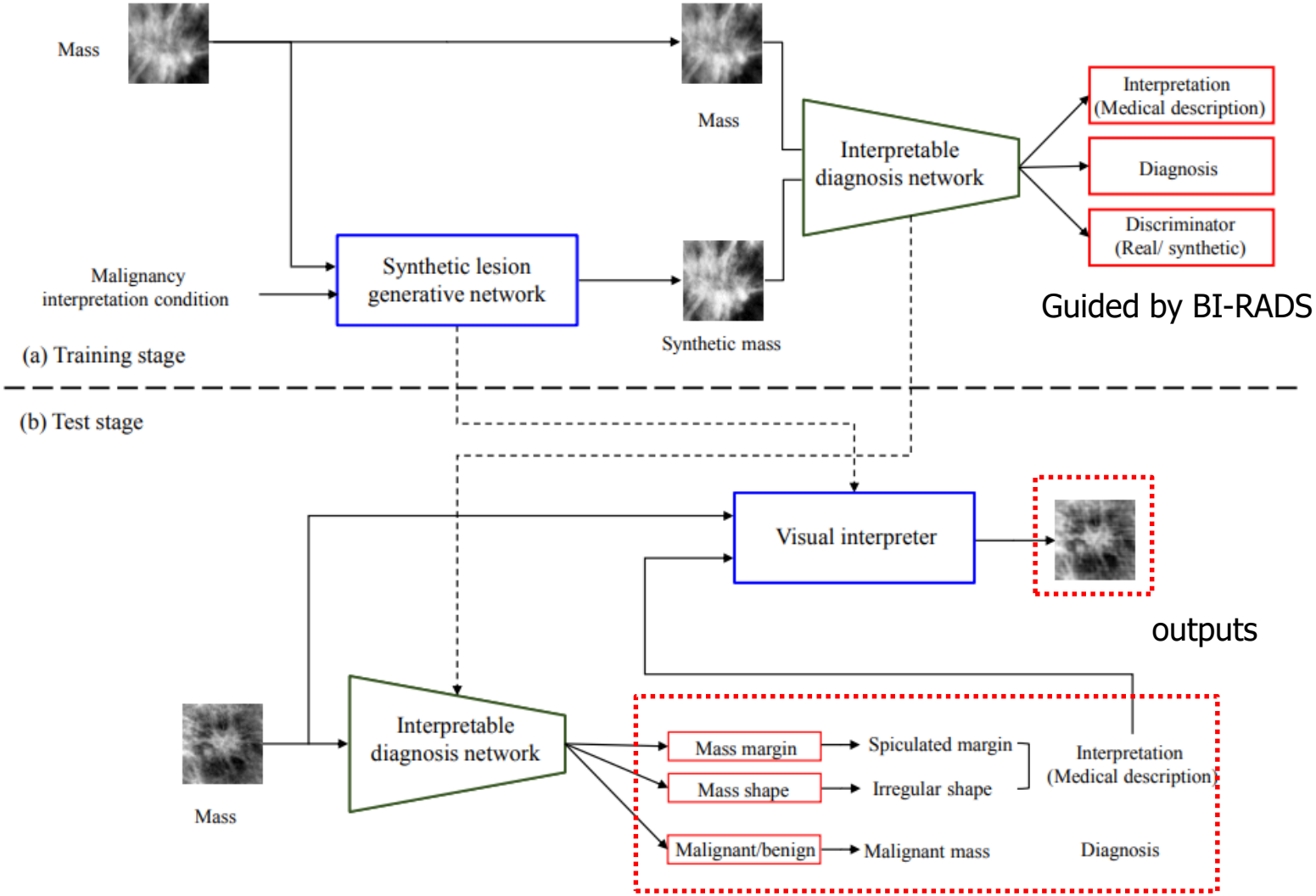
Explanations that are familiar to doctors

**BI-RADS: Breast Imaging Reporting and Data System*

Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro, "ICADx: Interpretable computer aided diagnosis of breast masses", SPIE Medical Imaging 2019 (Best Student Paper Award)

XCAD: Explainable Computer Aided Diagnosis – 2/2

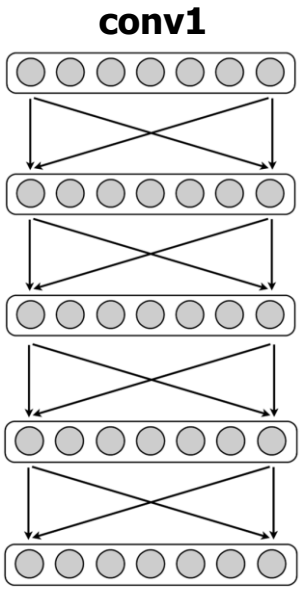
Yong Man Ro (KAIST)



Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro, "ICADx: Interpretable computer aided diagnosis of breast masses", SPIE Medical Imaging 2019 (Best Student Paper Award)

Network Dissection

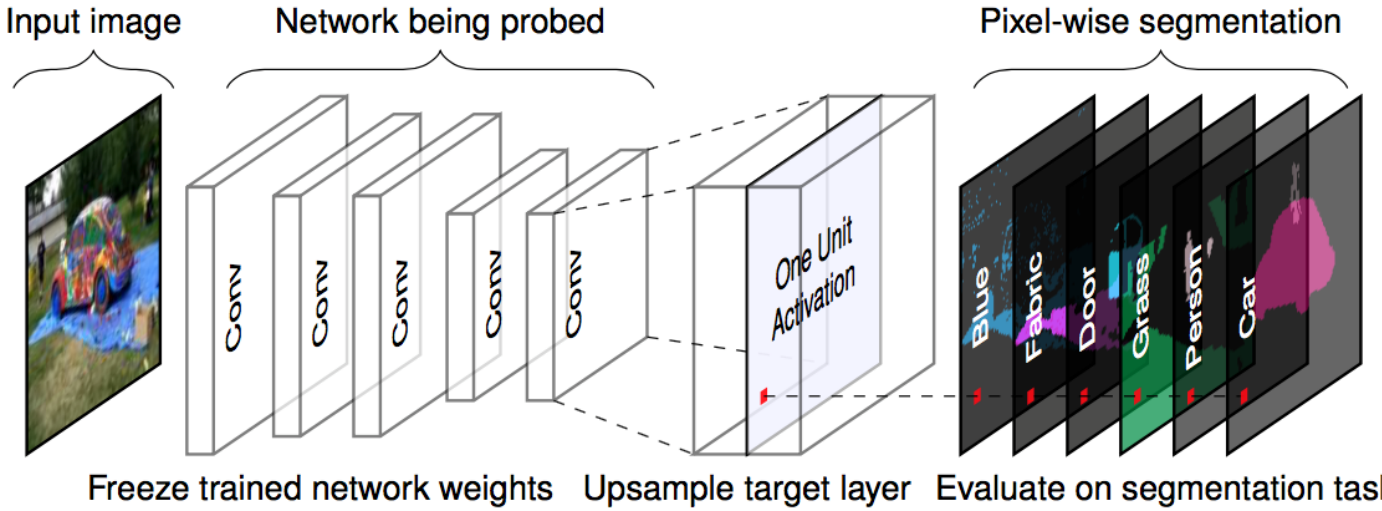
David Bau et. al., 2017 (A. Torralba, MIT)



Concepts Recognized in AlexNet Layers for Recognizing Places

Network Dissection

David Bau et. al., 2017 (A. Torralba, MIT)



Network Dissection

David Bau et. al., 2017 (A. Torralba, MIT)

Buildings

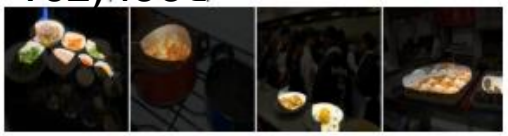
56) building

8) bridge

123) building

Indoor objects

182) food



46) painting

106) screen

Furniture

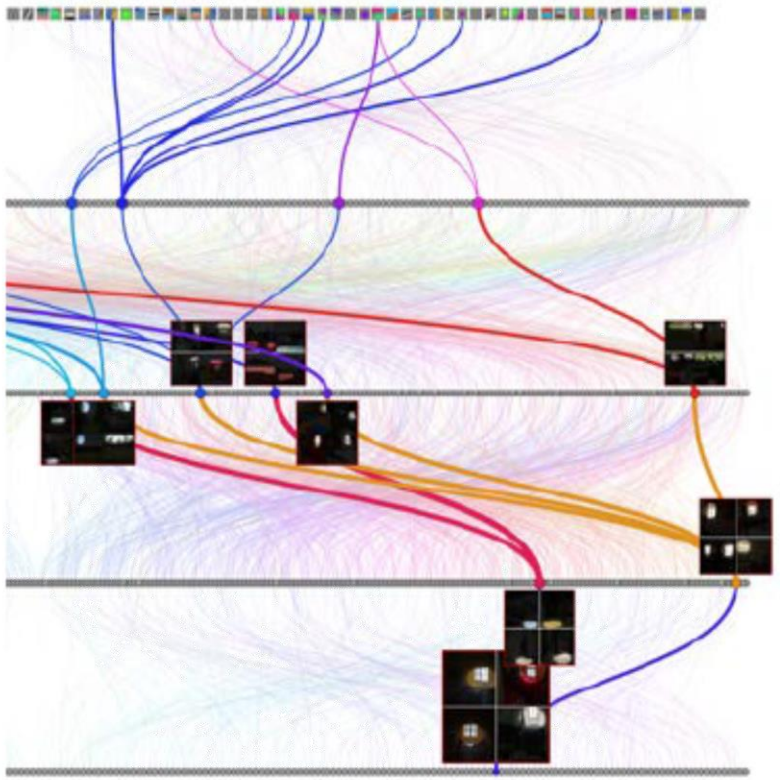
18) Billiard table

155) bookcase

116) bed

Interpretation of several units in pool5 of AlexNet trained for place recognition

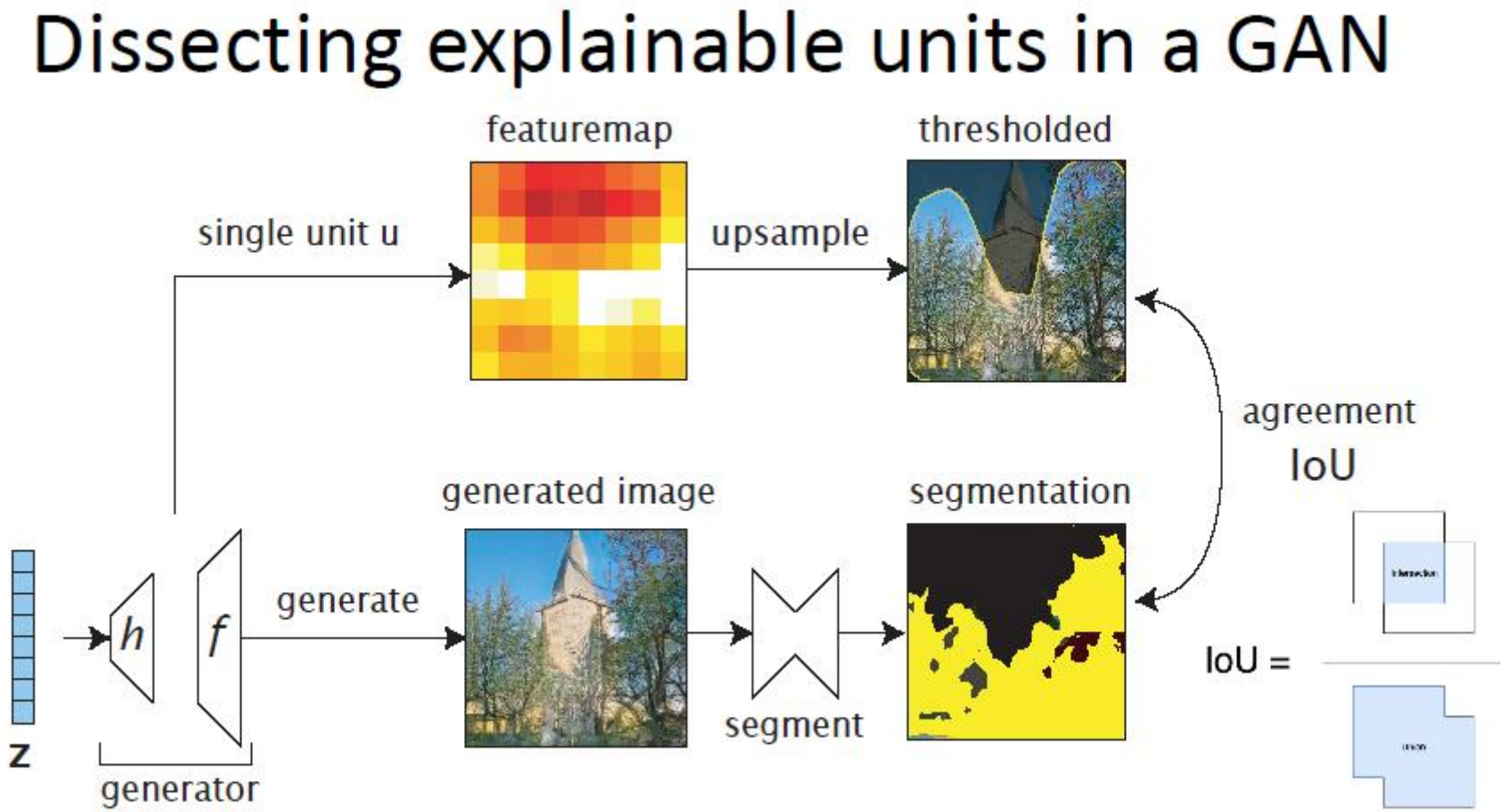
Audit trail: for a particular output unit, the drawing shows the most strongly activated path



A. Torralba (MIT) et. al., 2017

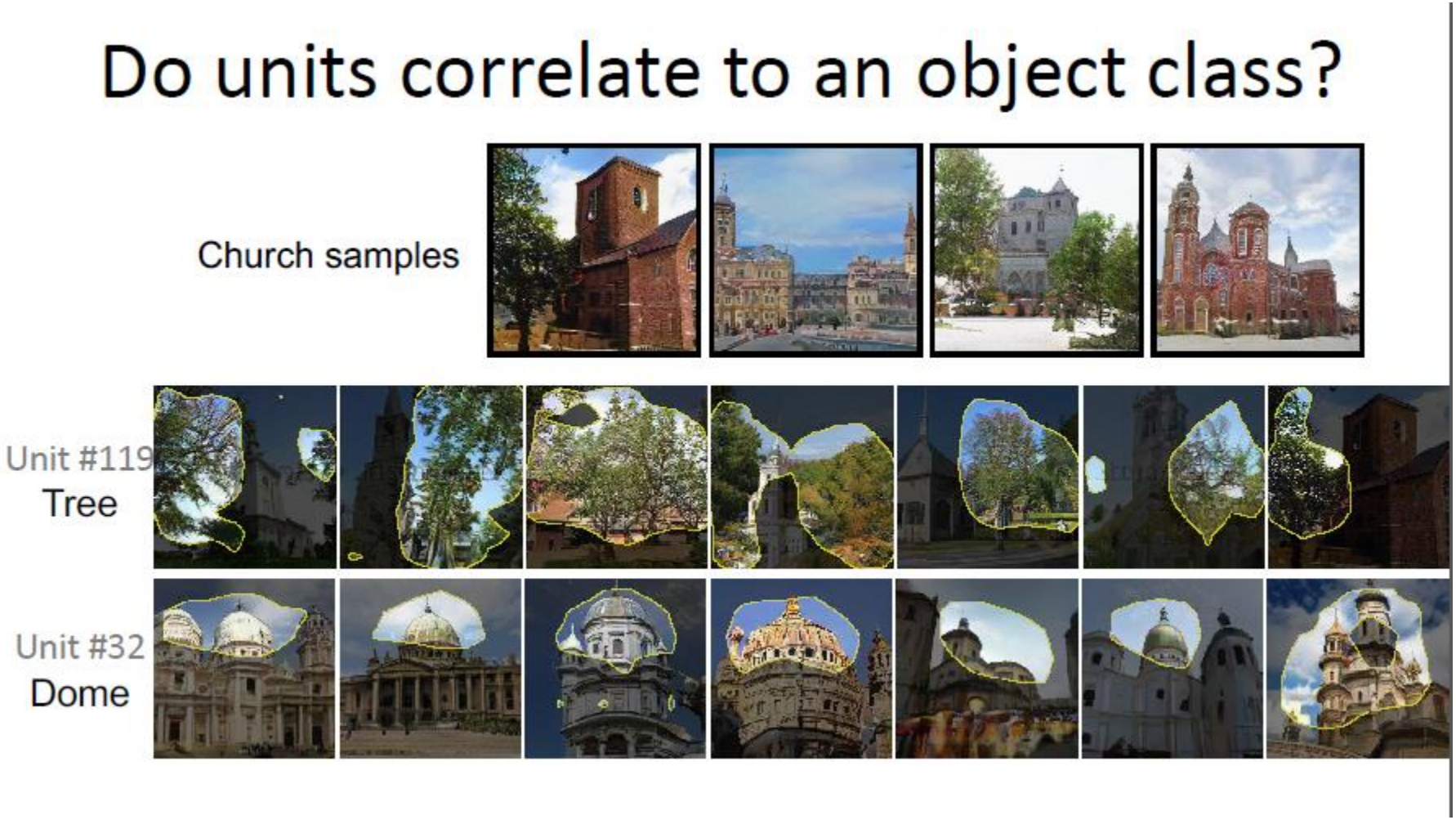
GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)



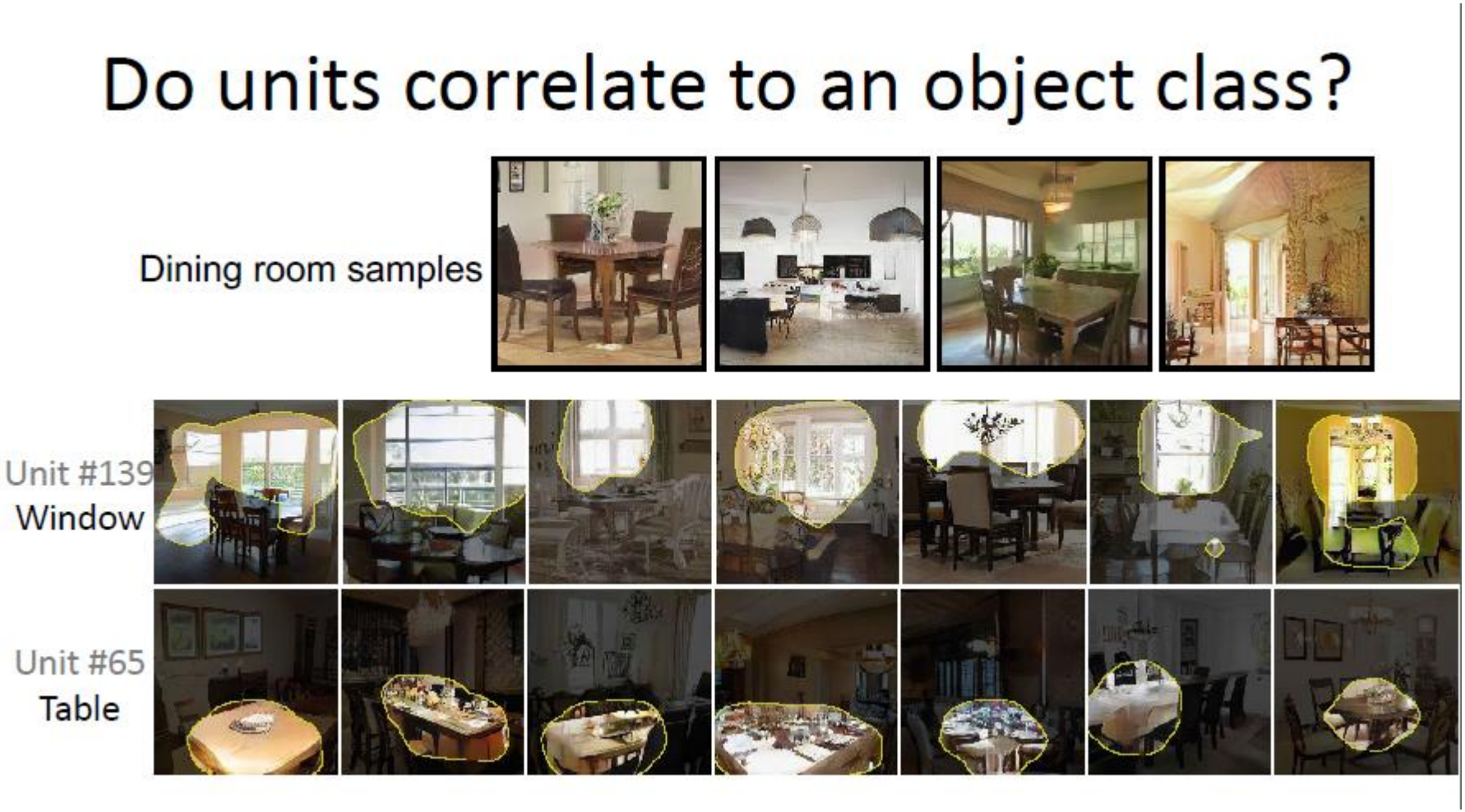
GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)



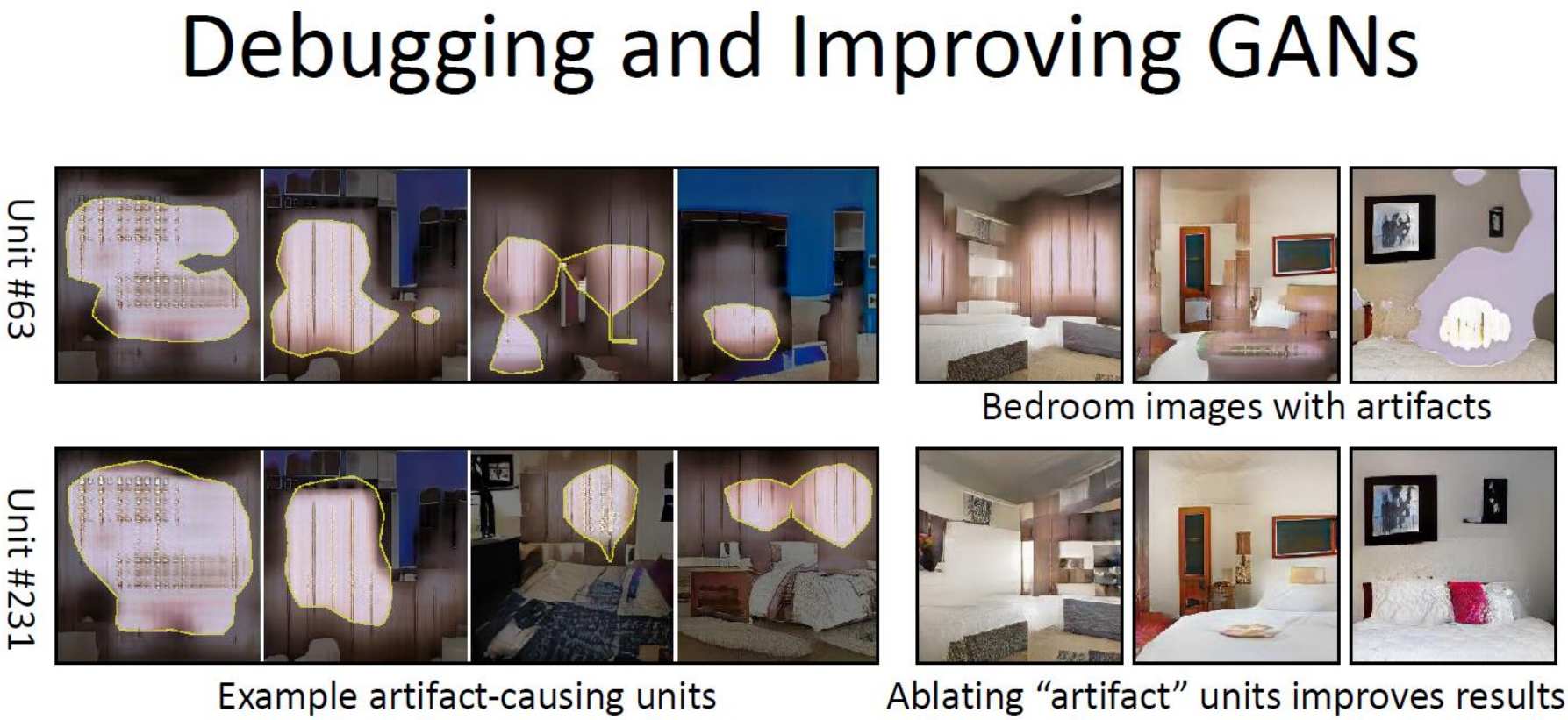
GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)



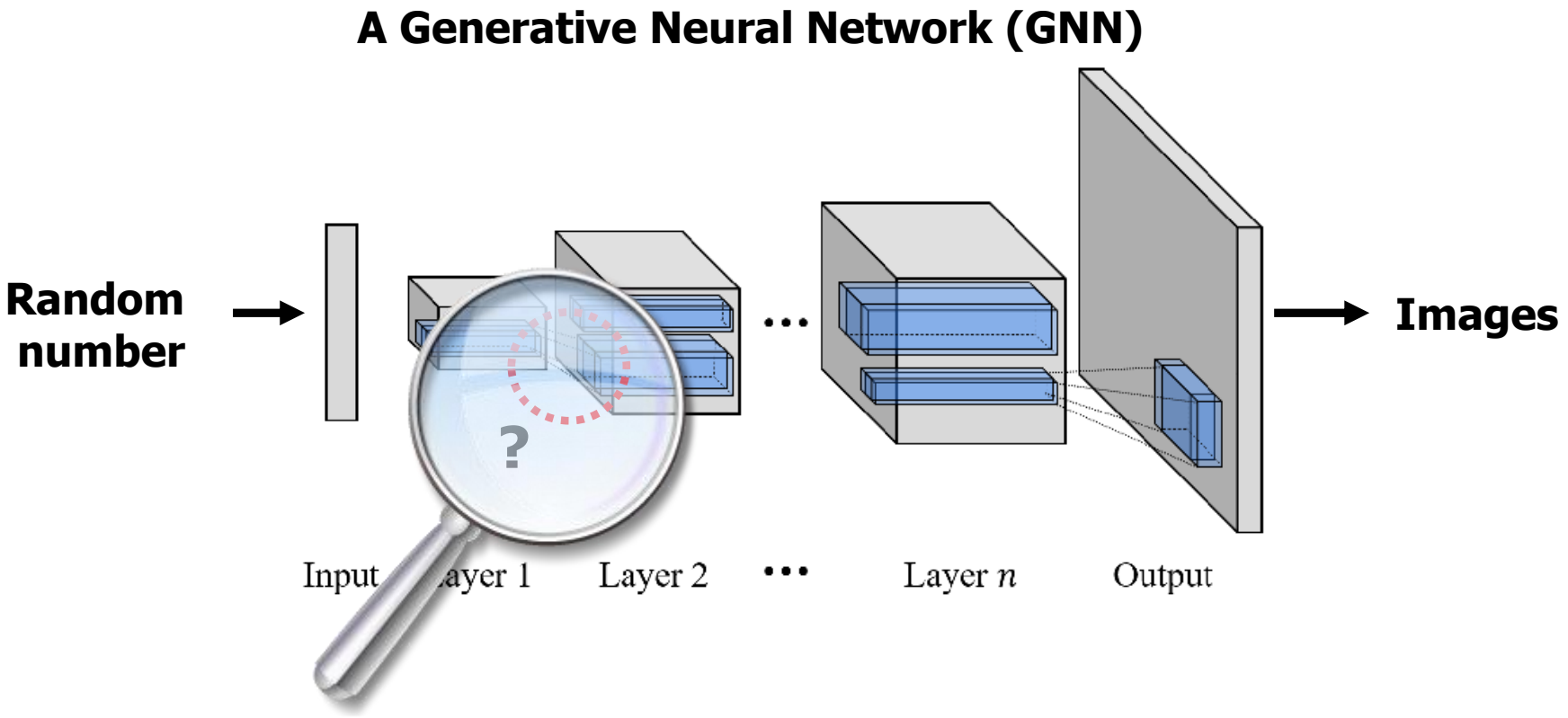
GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)



Generative Boundary Aware Sampling

Giyoung Jeon et. al., 2019 (J. Choi, UNIST)

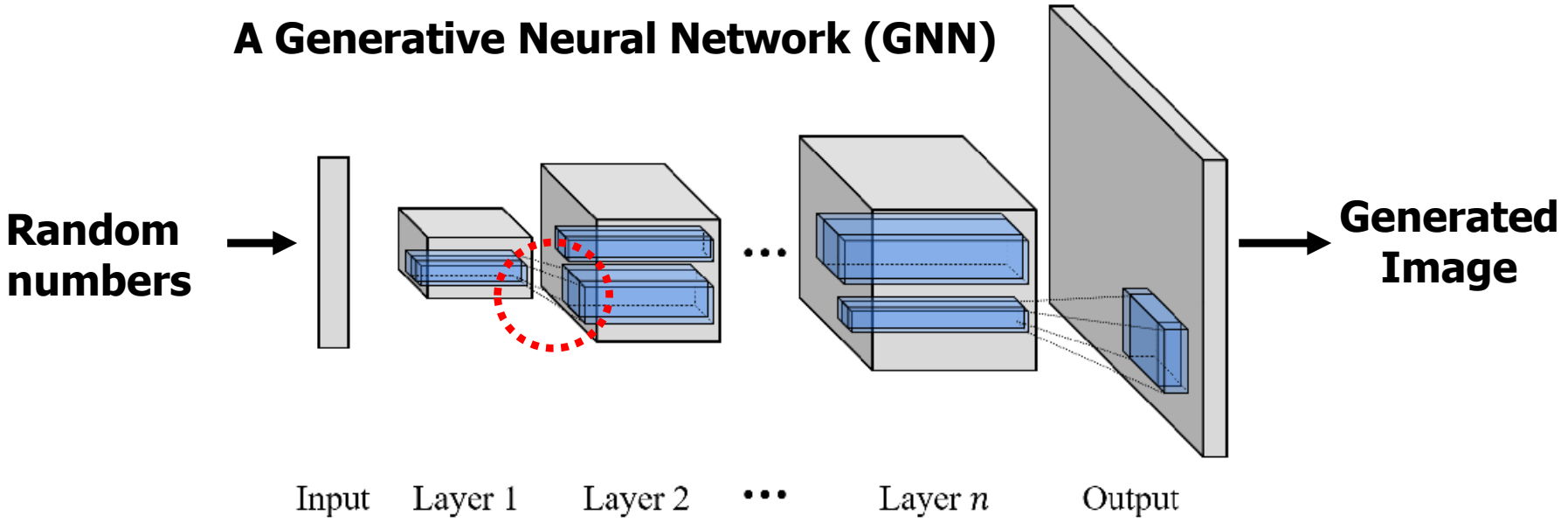


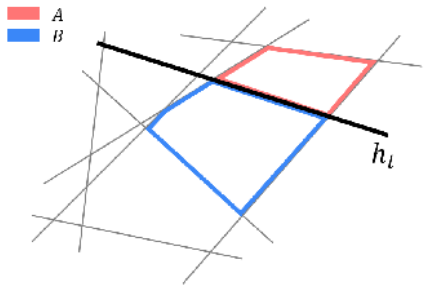
Dissecting Deep Neural Networks

Generative Boundary Aware Sampling

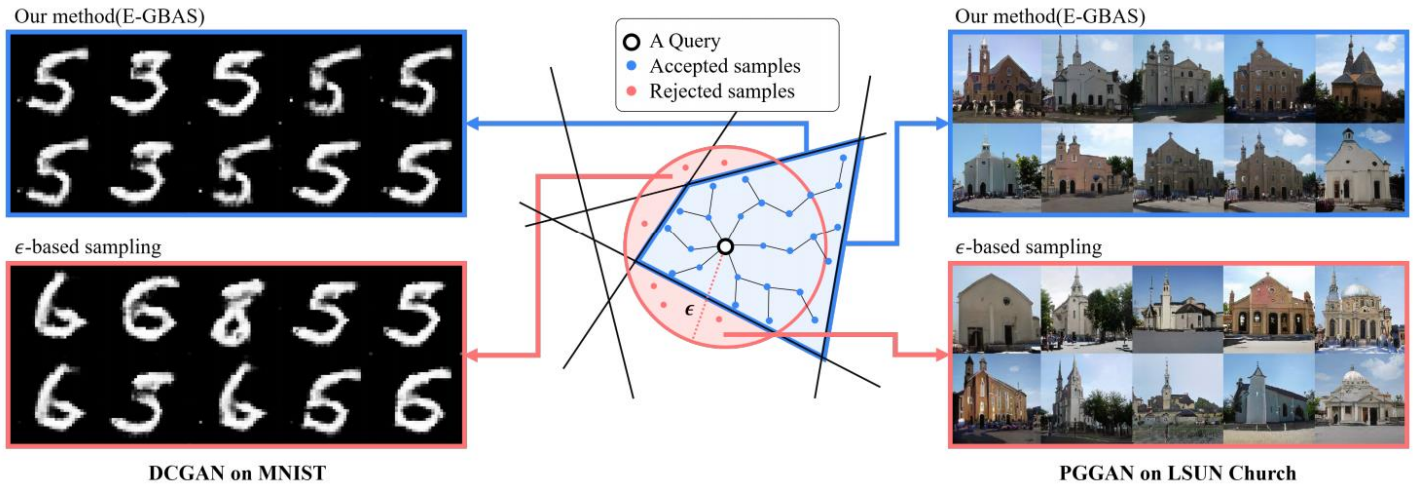
Giyoung Jeon et. al., 2019 (J. Choi, UNIST)

A Generative Neural Network (GNN)



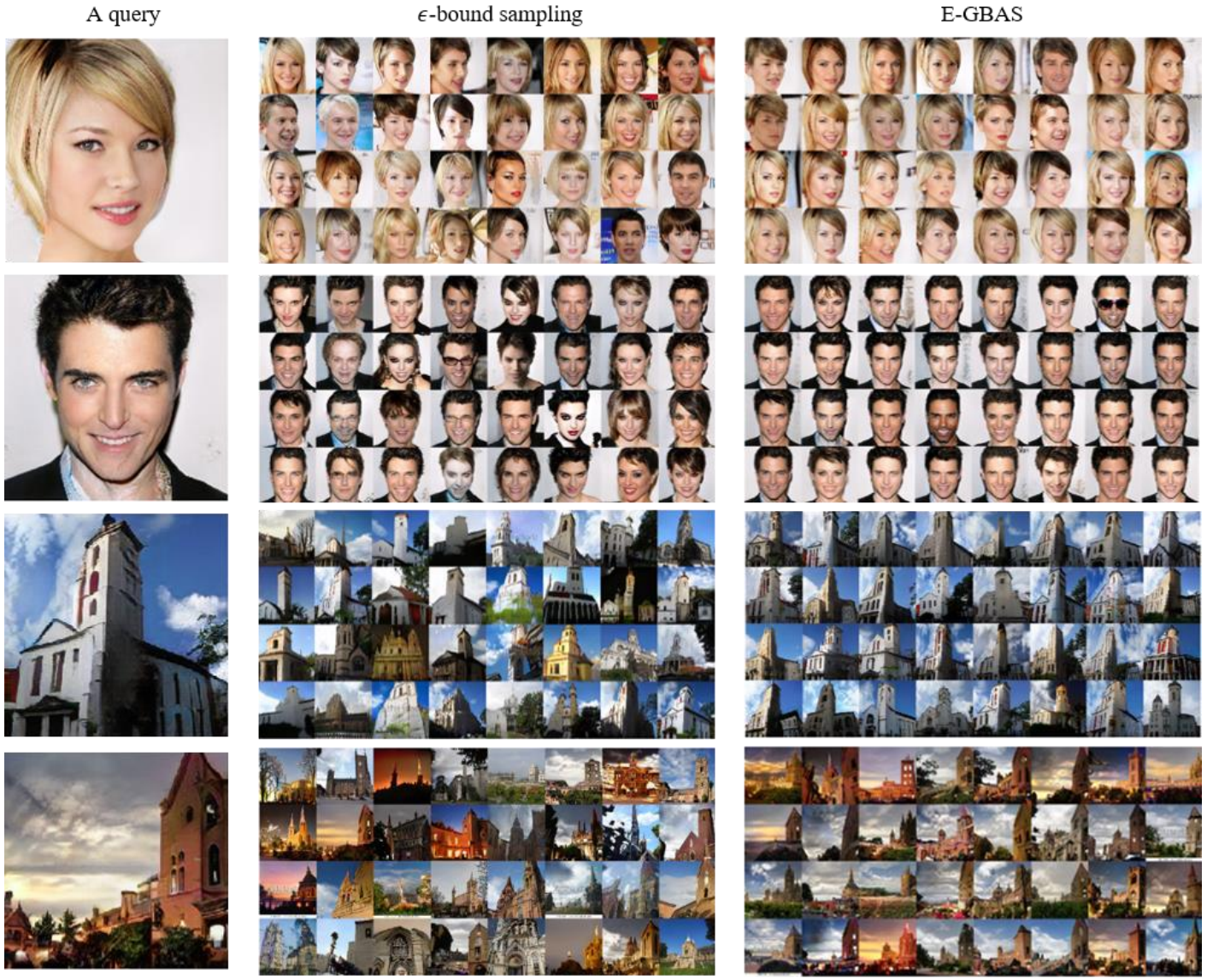


Generative boundaries



Generative Boundary Aware Sampling

Giyoung Jeon et. al., 2019 (J. Choi, UNIST)



Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

- **Extract State Representation of DRL (Oregon)**
- **Modular DRL (Berkeley)**

Explaining by Combining Explainable Models

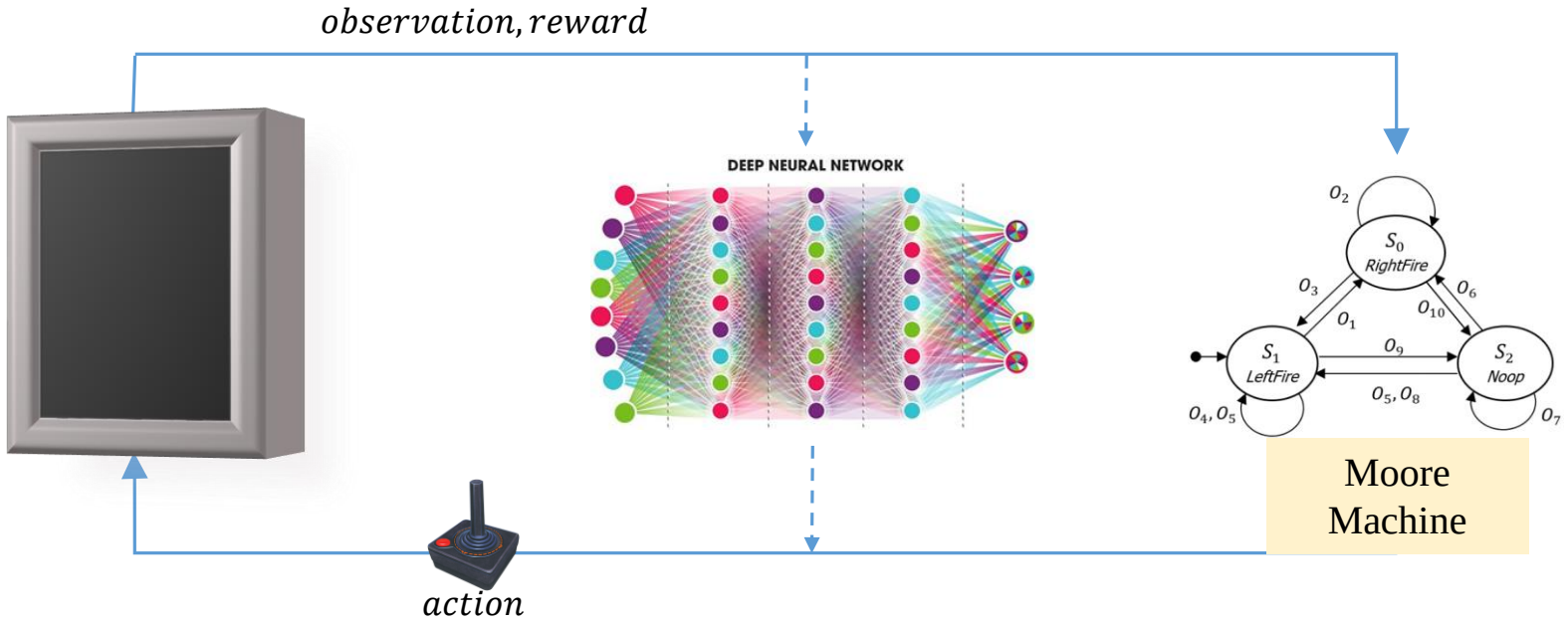
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Extract State Representation of DRL

Alan Fern (Oregon State)

Objective



Learning Finite State Representations of Recurrent Policy Networks
Anurag Koul, Alan Fern, and Sam Greydanus
International Conference on Learning Representations (ICLR-2019).

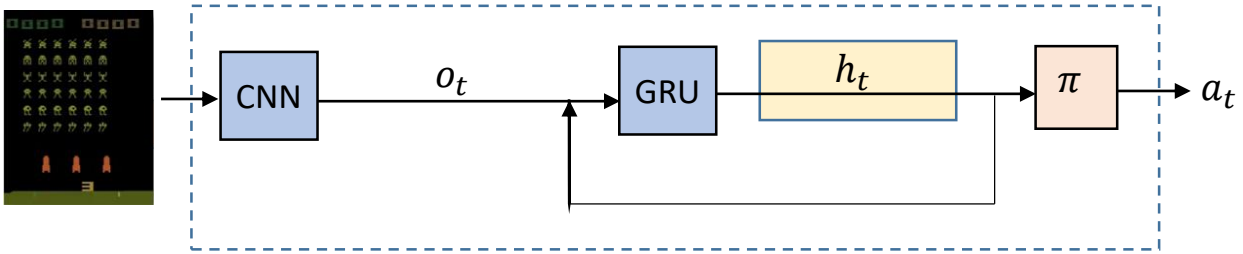
Oregon State University

Extract State Representation of DRL

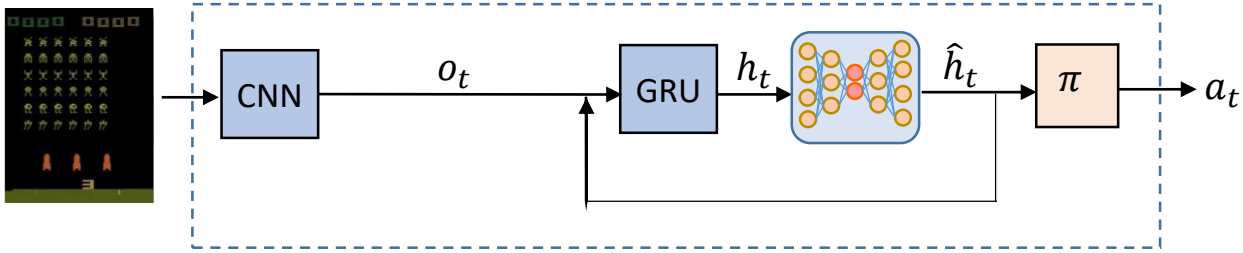
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

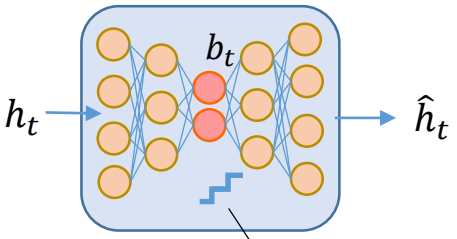
Original Continuous RNN



Discrete Memory



Quantized Bottleneck Network



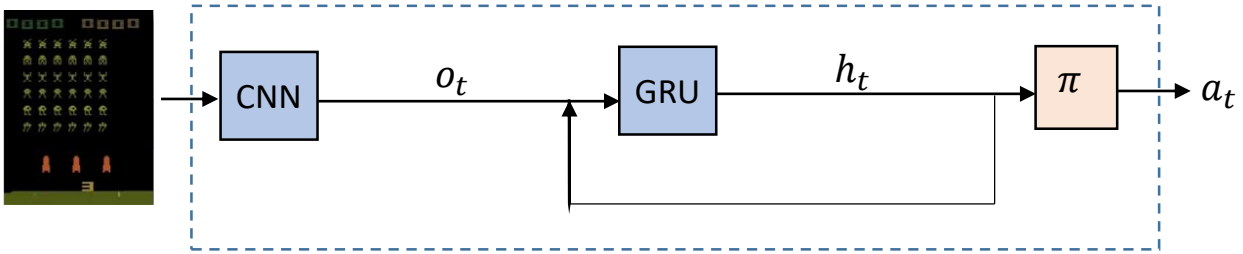
Oregon State University

Extract State Representation of DRL

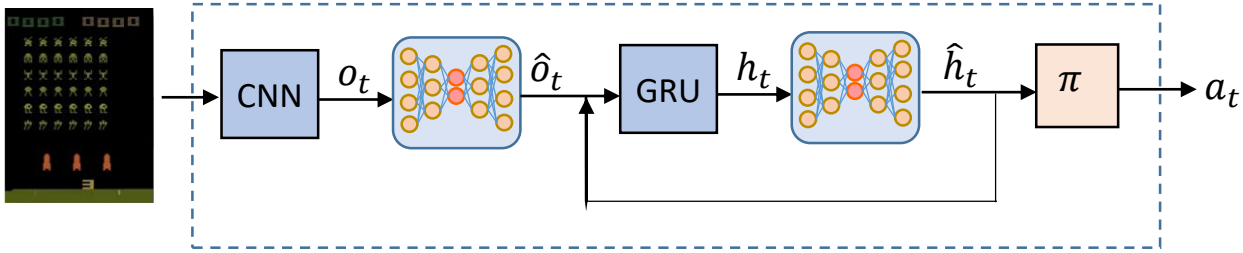
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

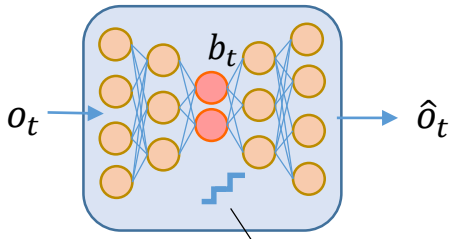
Original Continuous RNN



Discrete Memory + Discrete Input



Quantized Bottleneck Network



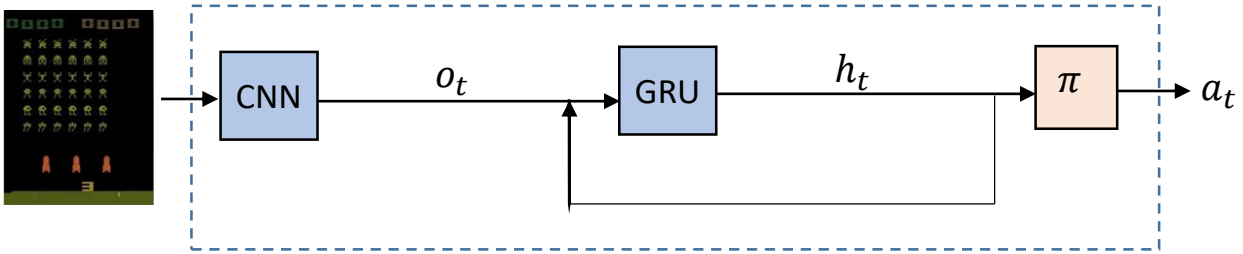
Discrete Representation

Extract State Representation of DRL

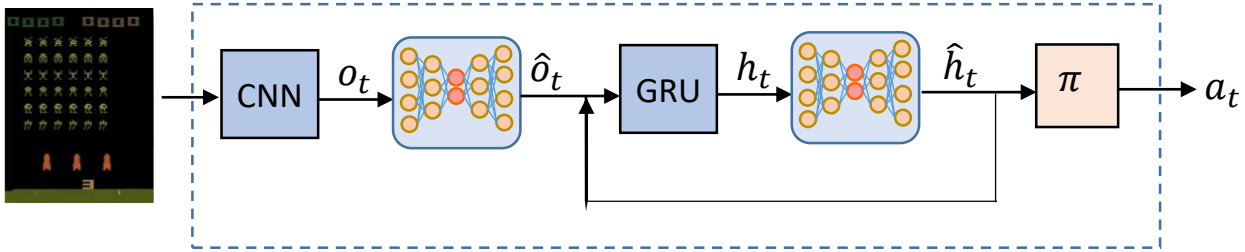
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

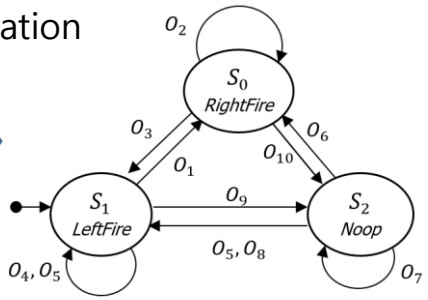
Original Continuous RNN



Discrete Memory + Discrete Input



Finite State Machine Minimization

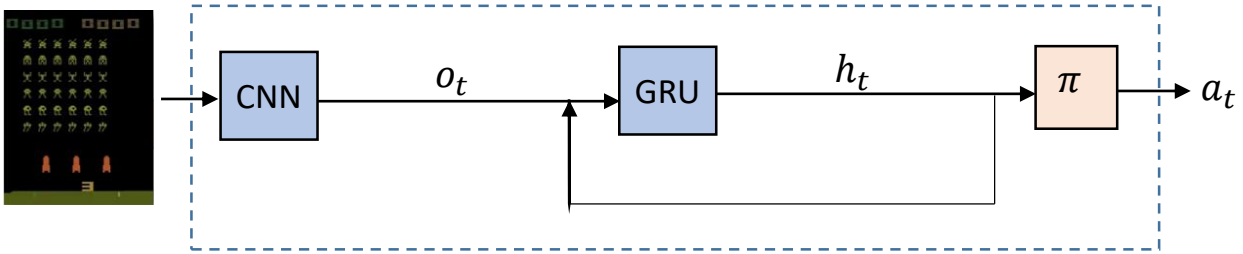


Extract State Representation of DRL

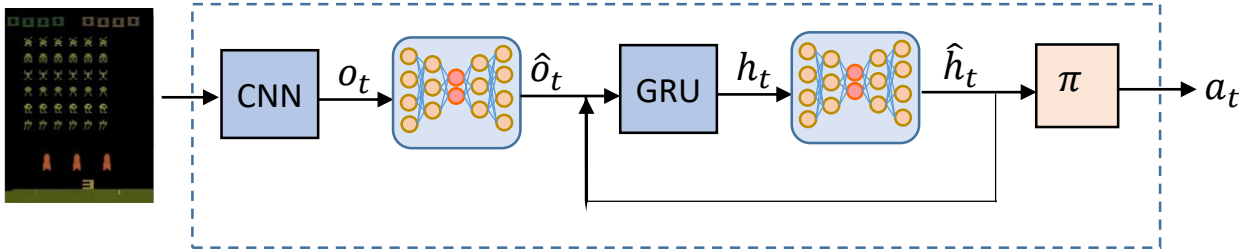
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

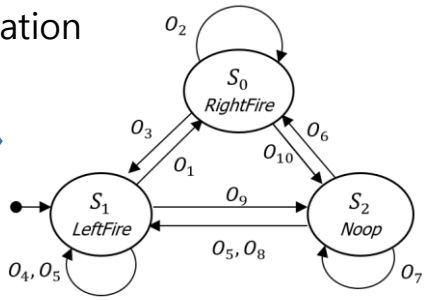
Original Continuous RNN



Discrete Memory + Discrete Input



Finite State Machine Minimization

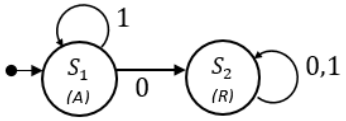


Extract State Representation of DRL

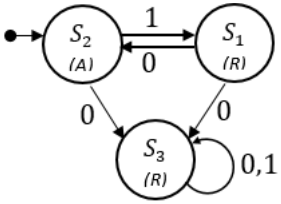
Alan Fern (Oregon State)

Grammar Learning Benchmarks: Tomita Grammars

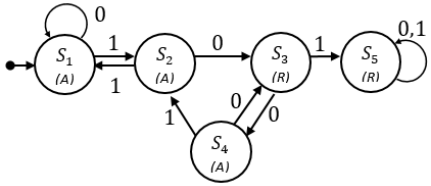
- *'A' and 'R' are acceptor and rejecter states*
- *Other than #6, all are 100% accurate. #6 is 99% accurate.*



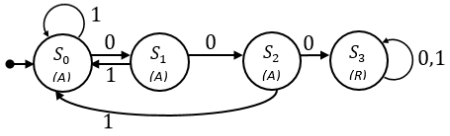
#1



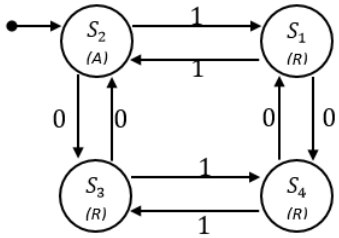
#2



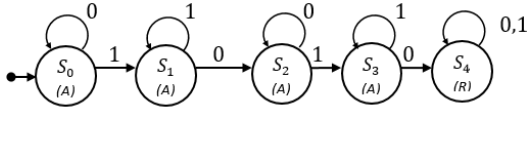
#3



#4

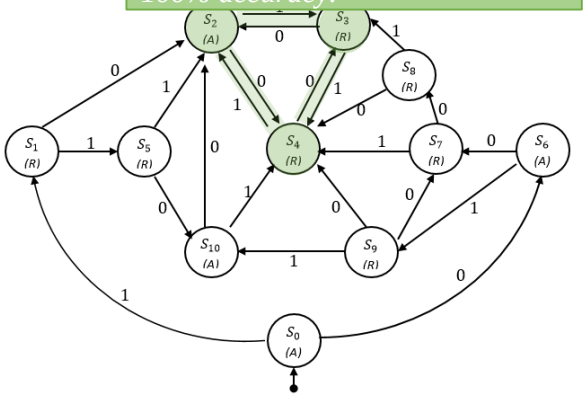


#5



#7

ground truth Moore machine for 100% accuracy.



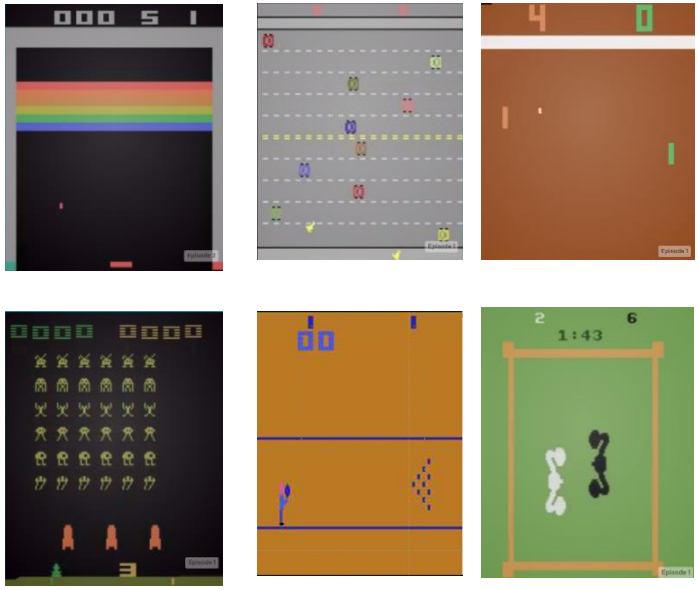
#6

Oregon State University

Extract State Representation of DRL

Alan Fern (Oregon State)

Atari – 6 Games



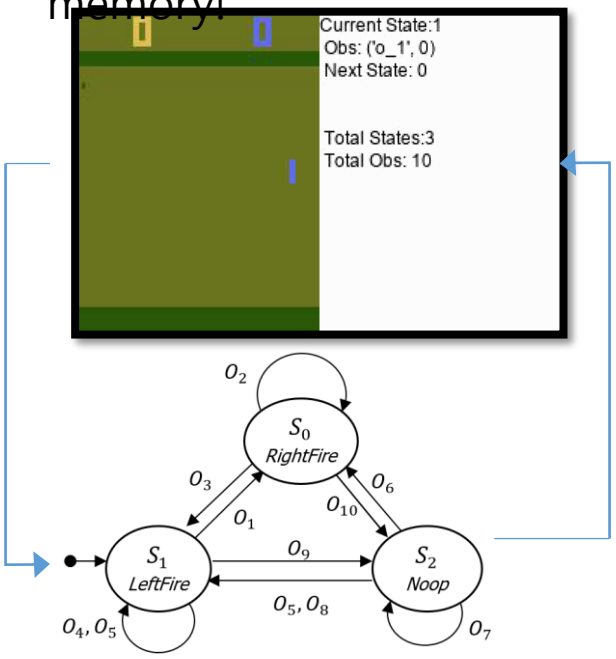
Game (# actions)	Stage - 4 Moore Machine					
	Before Minimization			**After Minimization**		
	H	**O**	Score	**H**	**O**	Score
Pong (3)	**373**	**372**	21	**3**	**10**	21
Freeway (3)	**1**	**1**	21	**1**	**1**	21
Breakout (4)	**1888**	**1871**	415	**8**	**30**	415
Space Invaders (4)	**1625**	**1620**	1235	**12**	**29**	1235
Bowling (6)	**49**	**1**	60	**33**	**1**	60
Boxing (18)	**2621**	**2605**	100	**14**	**119**	100

Extract State Representation of DRL

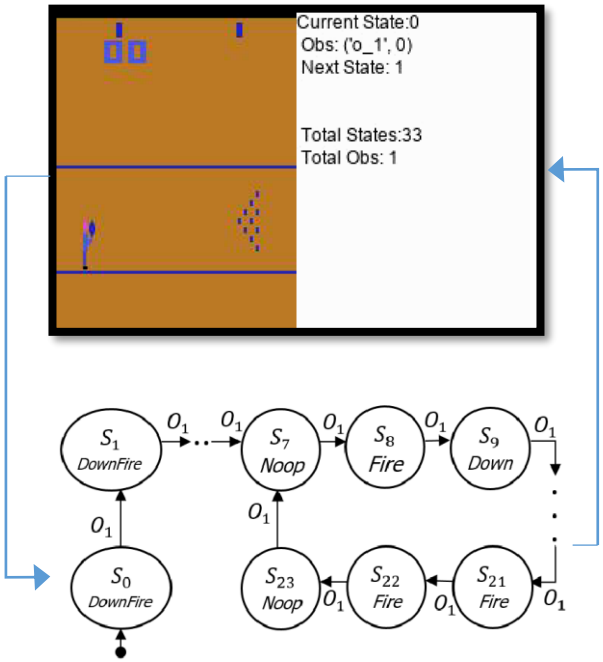
Alan Fern (Oregon State)

Atari – Demonstration (3 of our 6 Atari Games)

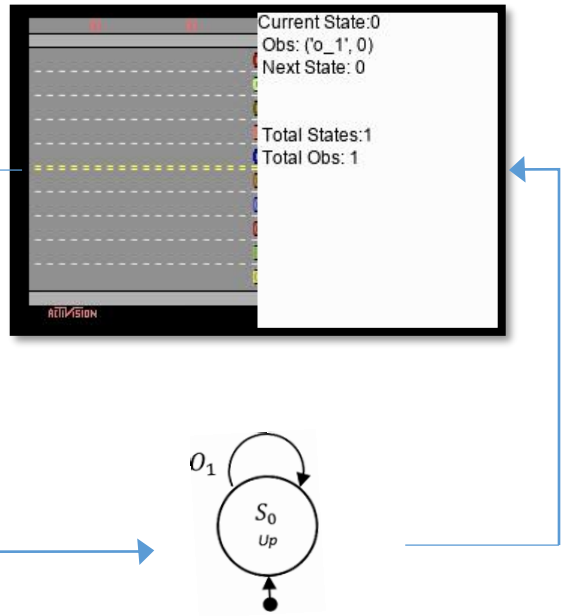
Pong
Policy does not use memory!



Bowling
Policy does not use observations!



Freeway
No memory nor observation



Modular DRL

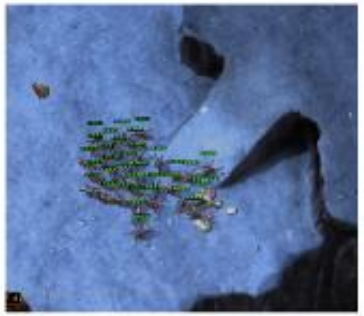
Pieter Abbeel (Berkeley)

- A popular real-time strategy game by Blizzard (2010 - now).
- Academic challenges: sparse rewards, large state & action spaces.
- Jan, 2019: DeepMind's AlphaStar beats pro players.

Collect Resources

Construct Buildings

Produce Units



Attack Enemies

[Dennis Lee*, Haoran Tang*, Jeffrey O Zhang, Huazhe Xu, Trevor Darrell, Pieter Abbeel AIIDE 2018]

Modular DRL

Pieter Abbeel (Berkeley)

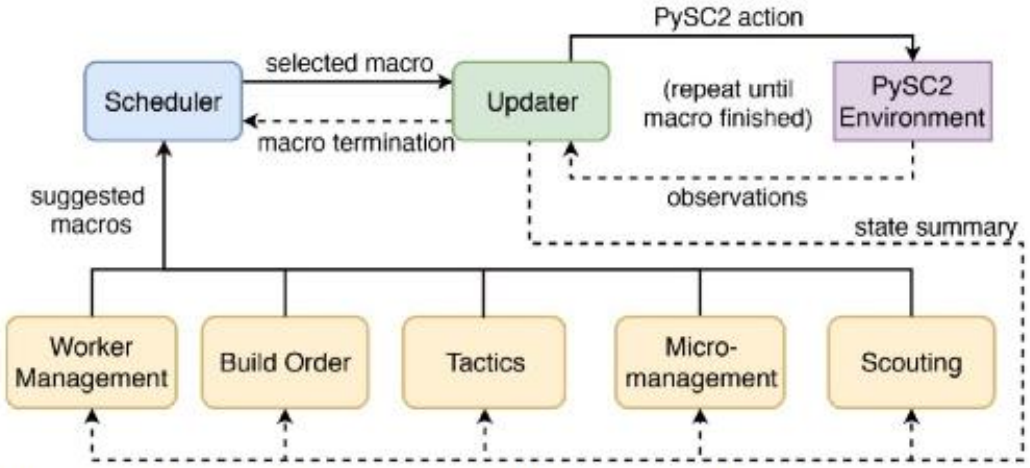
How to simplify the complex decision process?

- Split responsibilities between five modules
- Routine modules are scripted
- Hard-to-script modules (Build Order, Tactics) are trained with deep reinforcement learning + self-play

Benefits

- Efficient Learning
- Explainable: automatically explains its decisions in text. e.g.

(06:10) [Micro] Attack enemy armies



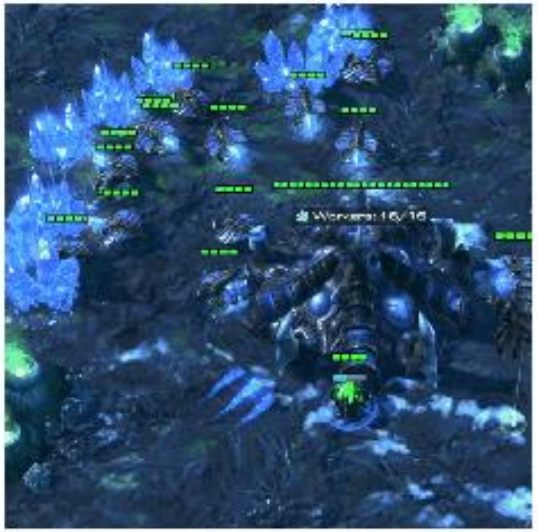
[Dennis Lee*, Haoran Tang*, Jeffrey O Zhang, Huazhe Xu, Trevor Darrell, Pieter Abbeel AIIDE 2018]

Modular DRL

Pieter Abbeel (Berkeley)

The experts are provided 3 examples of explaining certain behavior.

(0:53) [Build Order] Build the second base



(01:24) [Worker Management] Move 3 workers to mine gas

(06:10) [Micro] Attack enemy armies

Modular DRL

Pieter Abbeel (Berkeley)

Without explanations

With explanations

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Explaining by Combining Explainable Models

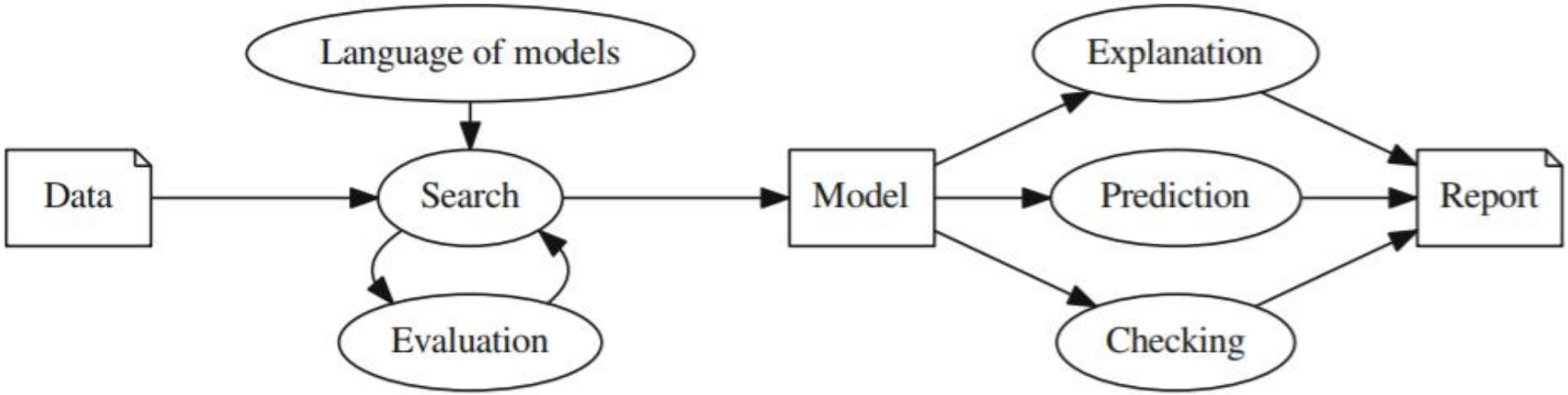
- **Automatic Statistician (MIT/Cambridge)**
- **Relational Automatic Statistician (UNIST)**

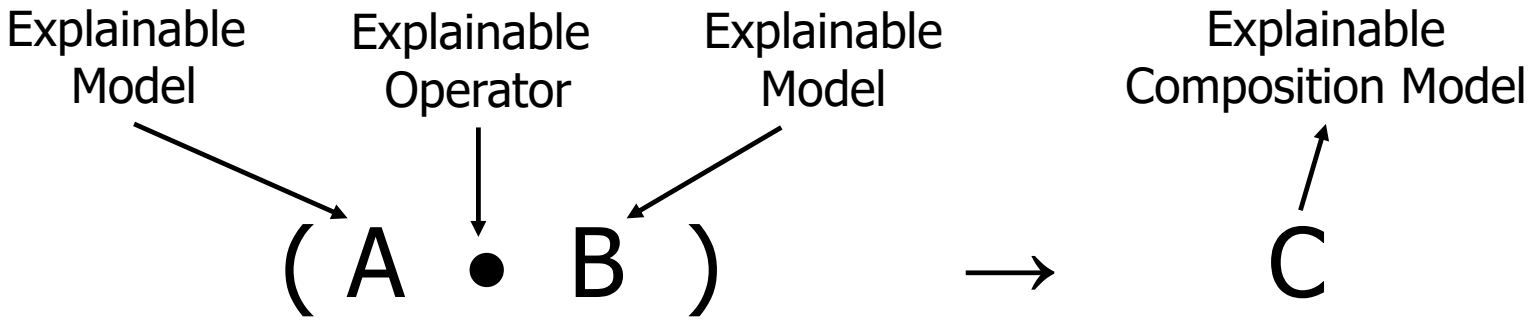
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Automatic Statistician

Z. Ghahramani (Cmabridge)
J. Tenenbaum (MIT)



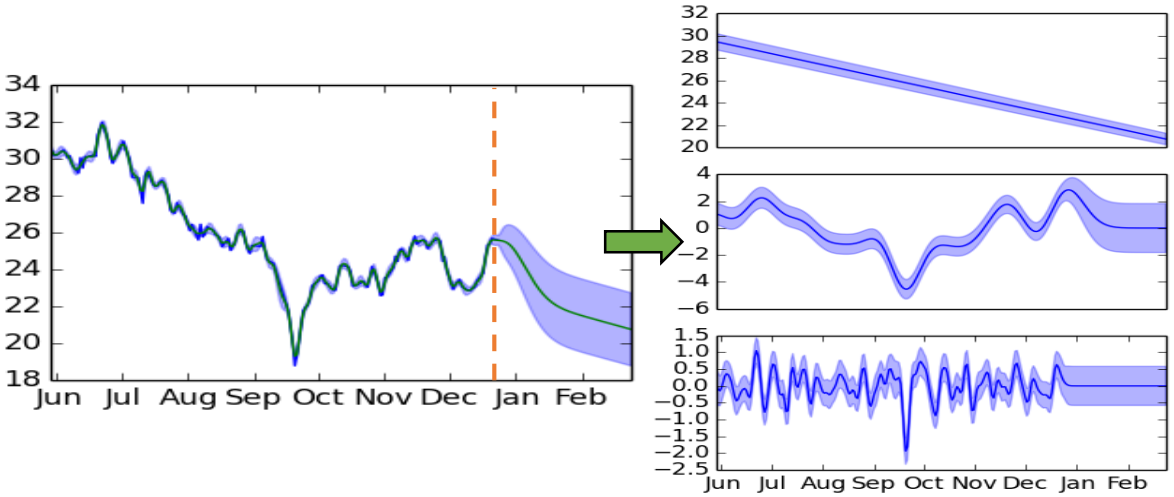


Automatic Statistician

Z. Ghahramani (Cmabridge)
J. Tenenbaum (MIT)

Explaining Multivariate Time Series in Financial Prediction

The Automatic Statistician System (Lloyd et al, 2014)

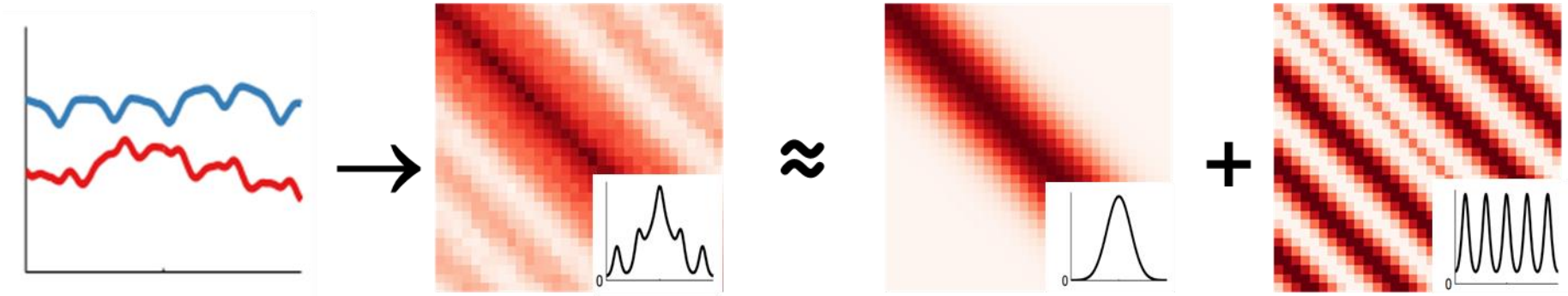


Linear function
decrease x/week

Smooth function
Length scale: y weeks

Rapidly varying
smooth function
Length scale: z hours

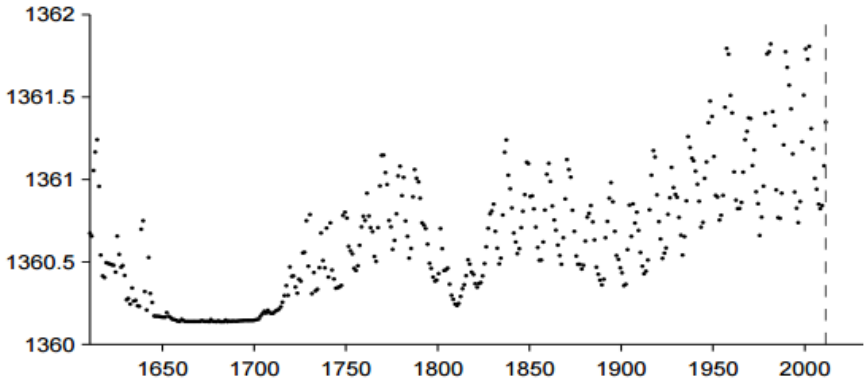
Covariance Decomposition: Learn Explainable Models from Data



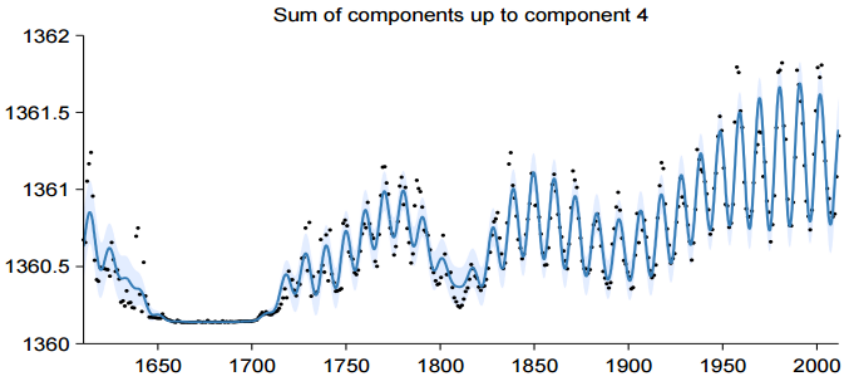
Automatic Statistician

Z. Ghahramani (Cmabridge)
J. Tenenbaum (MIT)

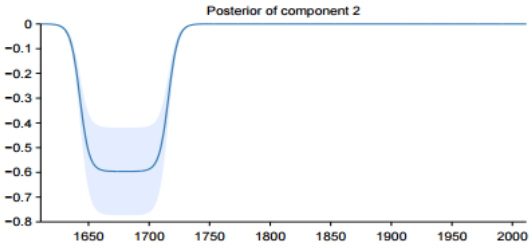
태양의 흑점 활동 데이터



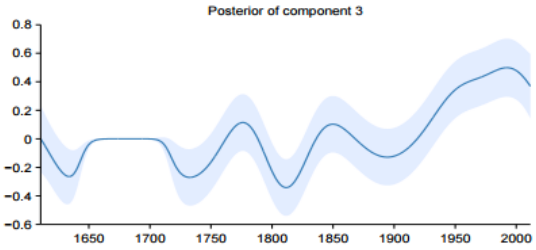
≈



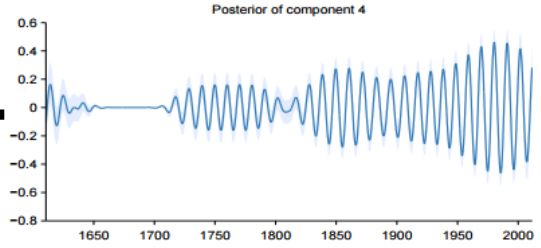
=



+



+

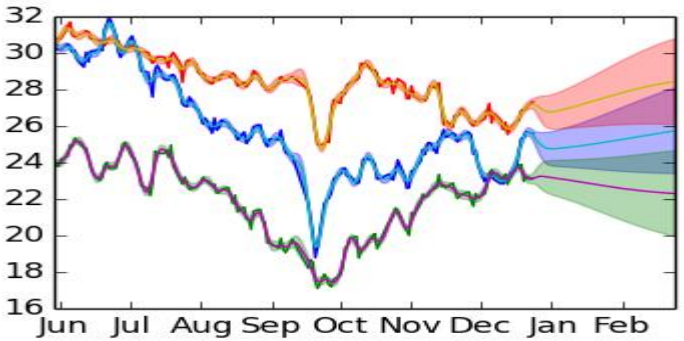


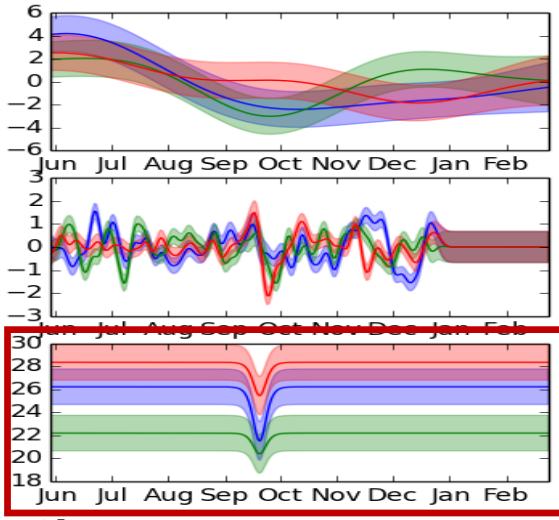
Relational Automatic Statistician

Jaesik Choi (UNIST)

Explaining Multivariate Time Series in Financial Prediction

Relational Automatic Statistician System (Hwang et al, ICML 2016)



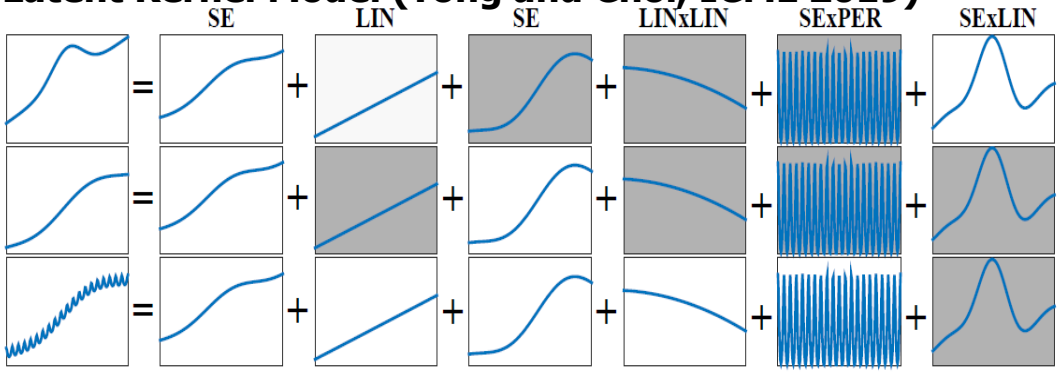


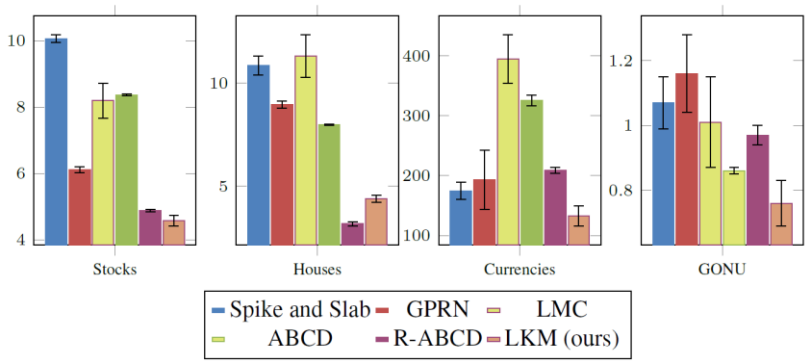
Smooth function
Length scale: y weeks

Rapidly varying
smooth function
Length scale: z hours

Constant function
Sudden drop btw
9/12/01 ~ 9/15/01

Latent Kernel Model (Tong and Choi, ICML 2019)





Comparisons of RMSE

Yunseong Hwang, Anh Tong and Jaesik Choi, Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series, ICML 2016

Anh Tong and Jaesik Choi, "Discovering Relational Covariance Structures for Explaining Multiple Time Series", ICML 2019

Relational Automatic Statistician

Jaesik Choi (UNIST)

Explaining Multivariate Time Series in Financial Prediction

An example of analysis of Gold, Oil, NASDAQ and USD

- Gold, Oil, NASDAQ, USD index share the following property:

This component is periodic with a period of 1.4 years but with varying amplitude. The amplitude of the function increases linearly away from Apr 2017. The shape of this function within each period has a typical lengthscale of 4.9 days.

- Gold, Oil, USD index share the following property:

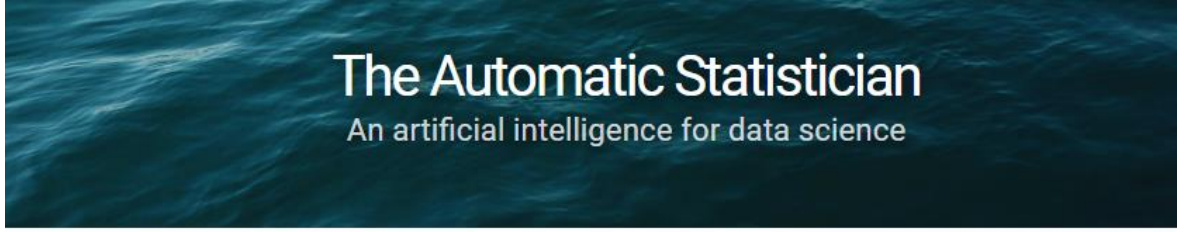
This component is a smooth function with a typical lengthscale of 2.7 weeks.

- NASDAQ has the following property:

This component is a linear function.

Yunseong Hwang, Anh Tong and Jaesik Choi, Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series, ICML 2016

Anh Tong and Jaesik Choi, "Discovering Relational Covariance Structures for Explaining Multiple Time Series", ICML 2019



http://www.automaticstatistician.com/

Research

Below is a selection of research articles relating to the Automatic Statistician project.

Scaling up the Automatic Statistician: Scalable Structure Discovery using Gaussian Processes
Hyunjik Kim, Yee Whye Teh
Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (2018)
pdf | supplementary pdf | arxiv | bibtex

The Automatic Statistician: A Relational Perspective
Yunseong Hwang, Anh Tong, Jaesik Choi
ICML 2016: Proceedings of the 33rd International Conference on Machine Learning (2016).
pdf | supplementary pdf | arxiv | slides | bibtex-1 | bibtex-2

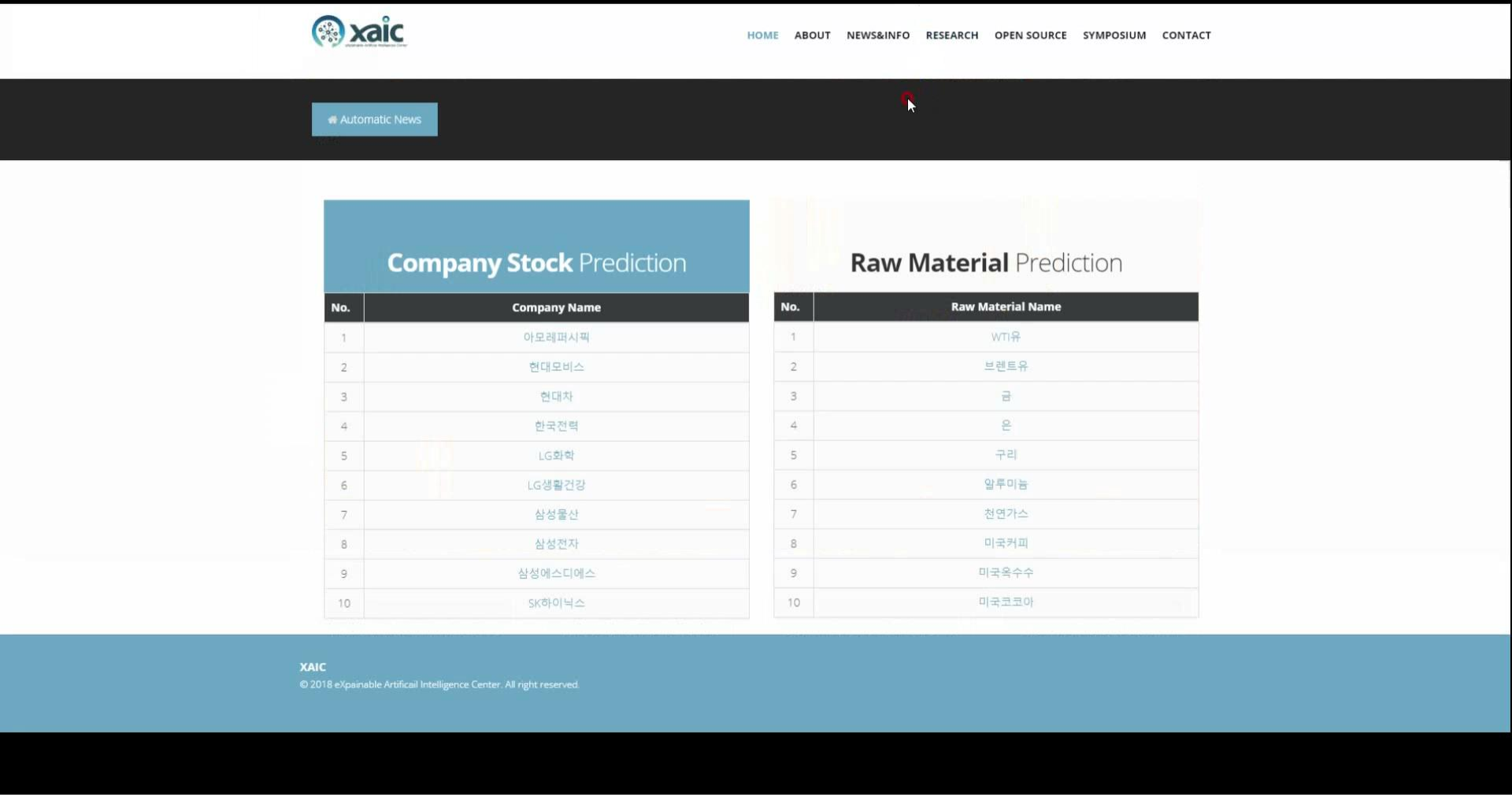
Statistical Model Criticism using Kernel Two Sample Tests
James Robert Lloyd, Zoubin Ghahramani
Advances in Neural Information Processing Systems 28 (2015)
pdf | code | bibtex

Automatic Construction and Natural-Language Description of Nonparametric Regression Models
James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, Zoubin Ghahramani
Association for the Advancement of Artificial Intelligence (AAAI) Conference, 2014
pdf | code | examples | bibtex

관계형 자동 통계학자(UNIST)
The Relational Automatic Statistician

Explaining Financial Prediction - Demo

Jaesik Choi & Byunggi Seo (UNIST)



Yunseong Hwang, Anh Tong and Jaesik Choi, Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series, ICML 2016

Anh Tong and Jaesik Choi, "Discovering Relational Covariance Structures for Explaining Multiple Time Series", ICML 2019

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Explaining by Combining Explainable Models

Finding Local Explanations (Model Agnostic Methods)

- **Counterfactual Generation (Toronto)**
- **SHapley Additive exPlanation (SHAP)**

Datasets and Applications

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

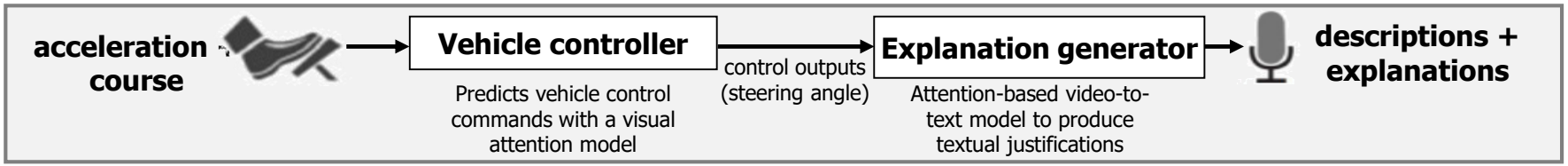
Explaining by Combining Explainable Models

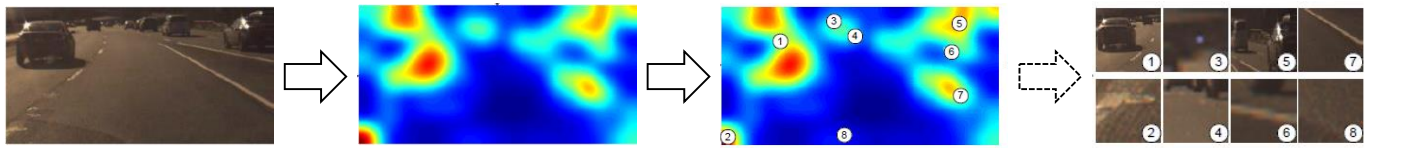
Finding Local Explanations (Model Agnostic Methods)

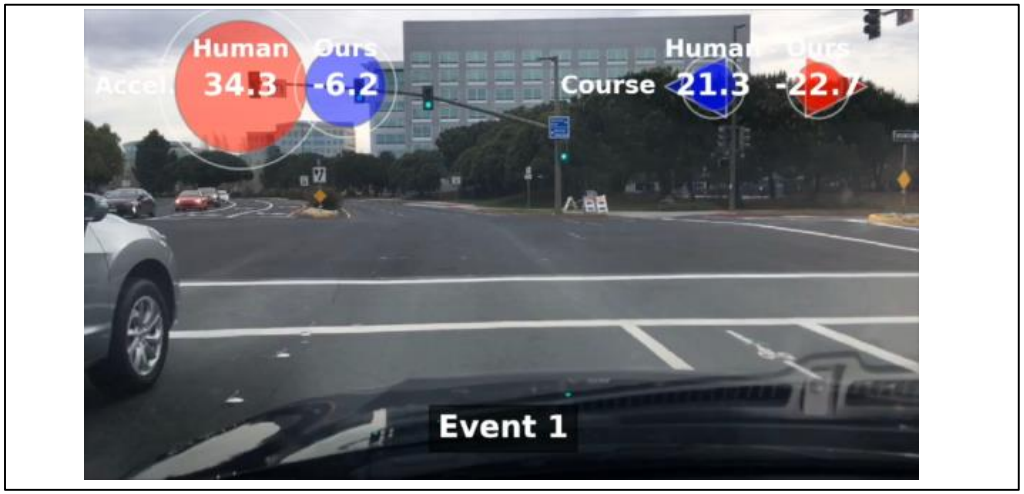
Data Sets and Applications

- **DeepDrive (Berkeley)**
- **Explaining Visual Causes (Yonsei)**
- **Distill (Google, OpenAI)**

UCB - Causal Grounded Driving

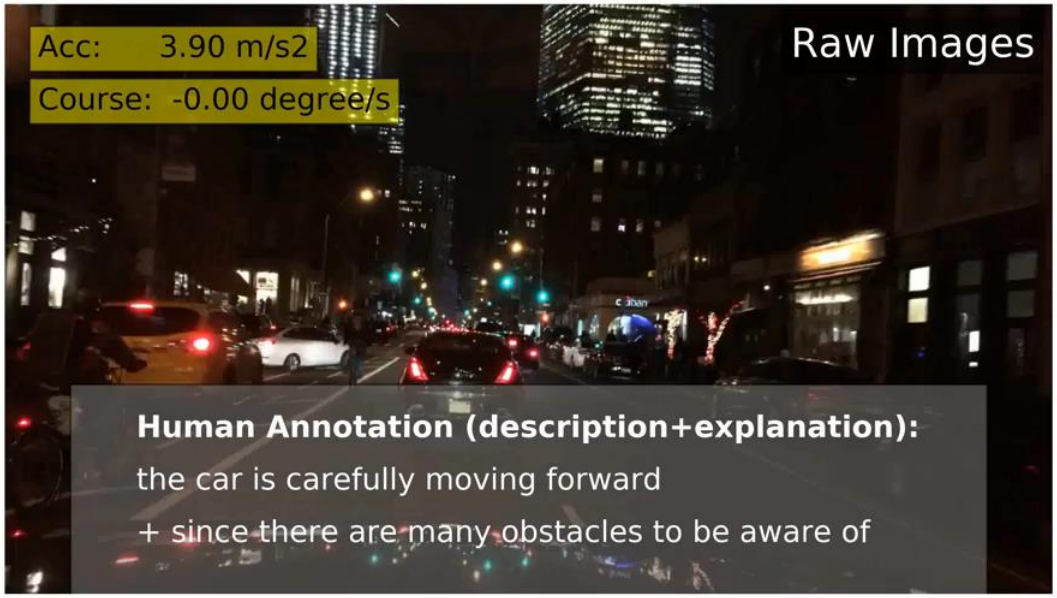






Kim and Canny, in ICCV, 2017
Kim, Rohrbach, Darrell, Canny, and Akata, in NIPS Interpretable ML Symposium, 2017

UCB - Causal Grounded Driving



Visual Choice of Plausible Alternatives

190 Dev set
190 Test set

Seung-won Hwang (Yonsei) **A new dataset for Visual Causal Reasoning** https://github.com/antest1/VCOPA-Dataset

(a) Visual disambiguation

(b) Temporal disambiguation

(c) Fine-grained object recognition

(d) Event recognition

(e) Inter-event relationship

(f) Event-sentiment relationship

(g) Inter-sentiment relationship

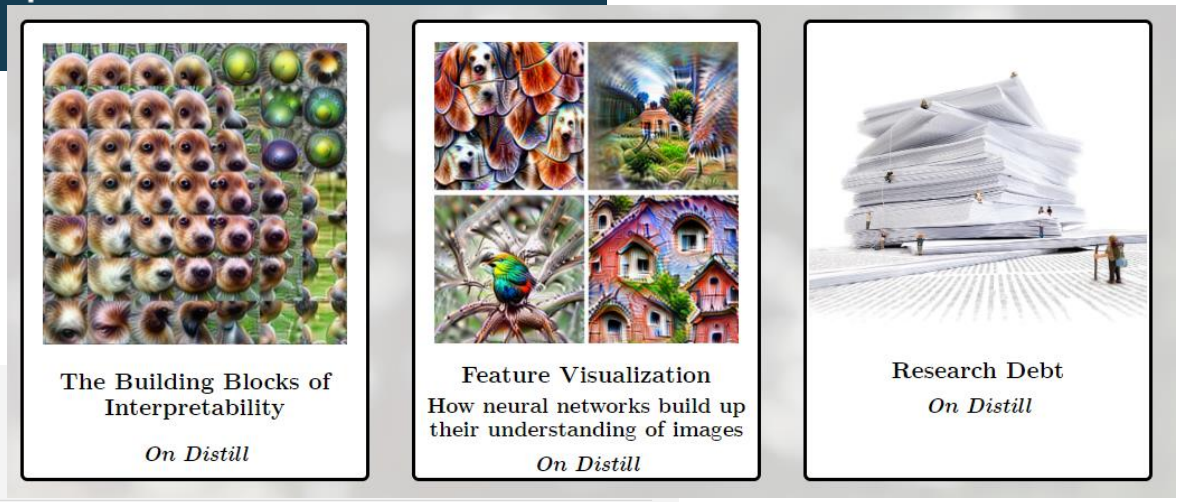
(h) Commonsense knowledge

Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo and Seung-won Hwang, Visual Choice of Plausible Alternatives: An Evaluation of Image-based Commonsense Causal Reasoning, LREC 2018

Distill for clear explanations of machine learning

OpenAI + Google

Machine Learning Research
Should Be Clear, Dynamic and Vivid.
Distill Is Here to Help.



The people behind Distill

EDITORS

Shan Carter
Google Brain Team

Chris Olah
OpenAI

Arvind Satyanarayan
MIT CSAIL

STEERING COMMITTEE

Yoshua Bengio
Université de Montréal

Mike Bostock
Data-Driven Documents

Amanda Cox
The New York Times

Ian Goodfellow
Google Brain Team

Andrej Karpathy
Tesla

Shakir Mohamed
DeepMind

Michael Nielsen
YC Research

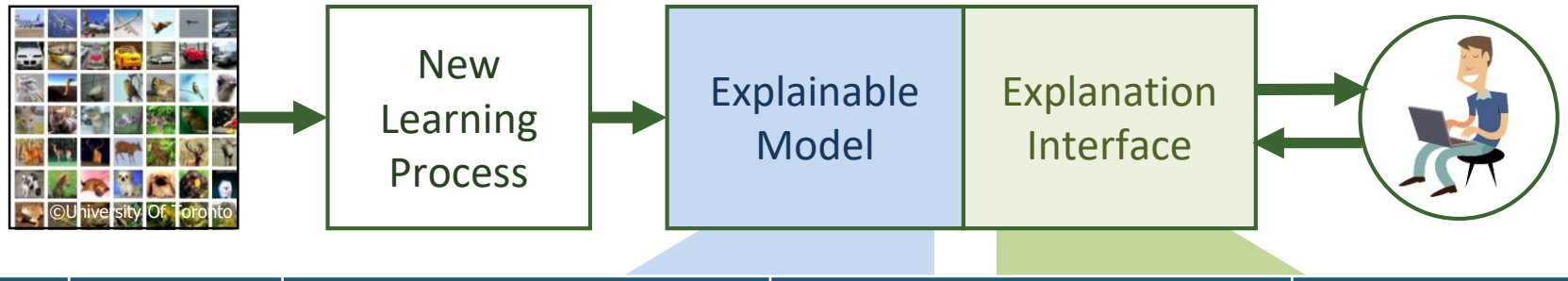
Fernanda Viégas
Google Big Picture

STAFF

Ludwig Schubert
OpenAI

Future Plan?

Future Plans



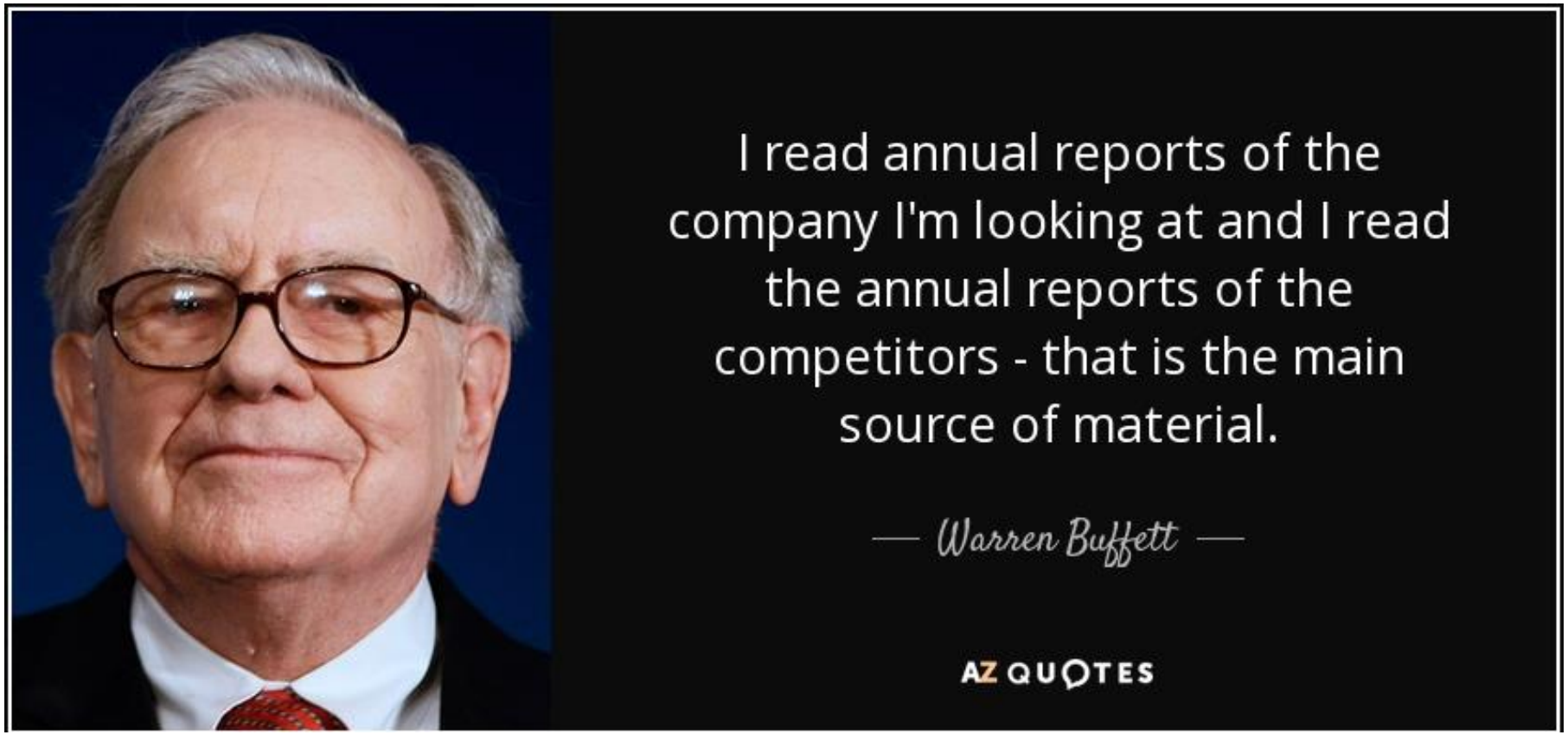
Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning	Textual Explanation	Visual Dialog
Task 3	UNIST	Nonparametric Bayesian	Narrative Explanation	Finance
	Korea*	Deep Learning (attribution methods)	Visual Explanation	Healthcare
	KAIST/ UNIST	Explainable RL	Interactive Visualization	Game

* Korea: Korea University

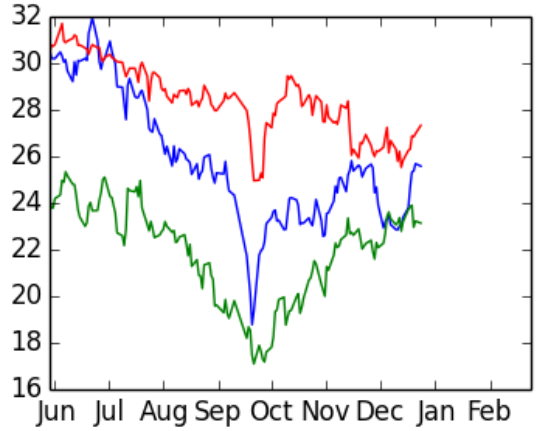
Future: Toward Reading/Explaining Reports

I read annual reports of the company I'm looking at and I read the annual reports of the competitors - that is the main source of material.

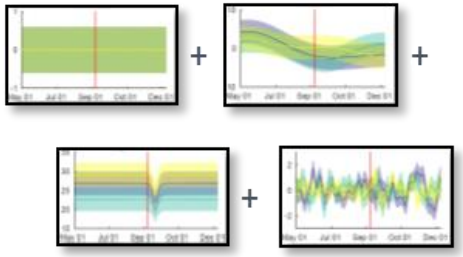
Future: Toward Reading/Explaining Reports



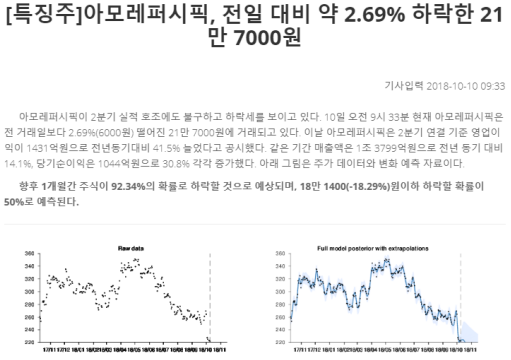
Future: Toward Reading/Explaining Reports



Multiple Stocks



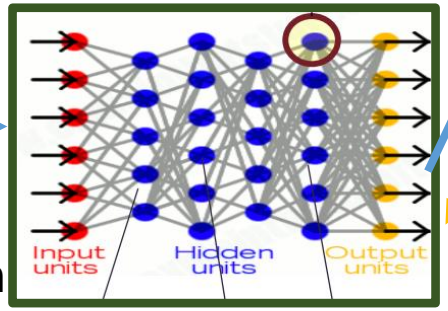
Bayesian Learning



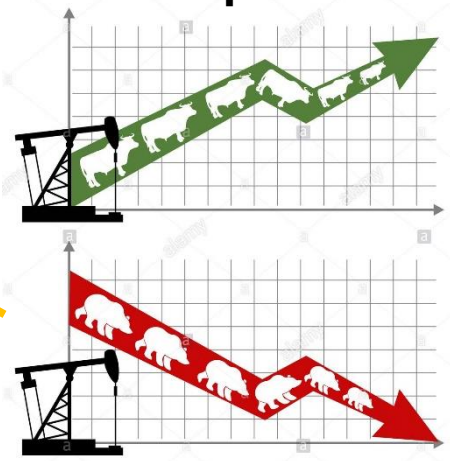
Report

Annual Reports

Find Explanation



Deep Learning



Prediction

Read the Reports and Data and Explain It

International Meetings

2018 International XAI Symposium

Trevor Darrell

Professor

UC Berkeley

Wojciech Samek

Head of Machine Learning Group

Fraunhofer Heinrich Hertz Institute

2019 ICCV Workshop on

Interpreting and Explaining Visual Artificial Intelligence Models

Saturday, November 2nd, 2019

@ COEX 318AB, Seoul, Korea

2019 International XAI Workshop

David Bau

PhD student

MIT

Ludwig Schubert

Software Engineer

OpenAI

Learn, Practice and Generate Knowledge to
Solve Some of the World's Greatest Problems in AI.

Thank you

jaesik@unist.ac.kr

Reference : site

- Part1
  - https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing
  - https://www.youtube.com/watch?v=7a6GrKqOxeU
  - https://www.youtube.com/watch?v=uHbMt6WDhQ8
  - https://www.youtube.com/watch?v=CrnLINlbfFA
  - https://www.youtube.com/watch?v=gdAVqJn_J2M

- Part2
  - https://www.imdb.com/interfaces/
  - https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714
  - https://wikidocs.net/32105
  - https://cyc1am3n.github.io/2018/11/10/classifying_korean_movie_review.html
  - http://xai.unist.ac.kr/opensource/relatedproject/

- Part3
  - Spring 2019 session (https://mlcourse.ai/tutorials) - ML interpretability by @Christophe Rigon

Reference : publication

- Marco Ancona, et. al., "Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation", ICML 2019
- Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018
- Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018
- Woo-Jeong Nam, Jaesik Choi, and Seong-Whan Lee, "Relative Attribution Propagation", arXiv 1904.00605, 2019
- Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro, "ICADx: Interpretable computer aided diagnosis of breast masses", SPIE Medical Imaging 2019 (Best Student Paper Award)
- David Bau et. al., "Network Dissection", CVPR, 2017 (A. Torralba, MIT)
- David Bau et. al., "GAN Dissection", 2019
- Kyowoon Lee, Sol-A Kim, Jaesik Choi, and Seong-Whan Lee, "Deep Reinforcement Learning in Continuous Action Spaces: a Case Study in the Game of Simulated Curling", ICML 2018
- Yunseong Hwang, Anh Tong and Jaesik Choi, "Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series", ICML, 2016
- Sebastian Bach et all, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation", ICANN, 2016

XAI센터의 연구결과를 더 알고싶다면

1. XAI 센터 공식홈페이지
http://xai.unist.ac.kr/

2. XAI 센터 깃허브
https://github.com/openxaiProject

3. XAI 센터 Youtube 채널
https://www.youtube.com/channel/UCGxsfIsOry_LdBaPSet2p7g

실습 환경 구축

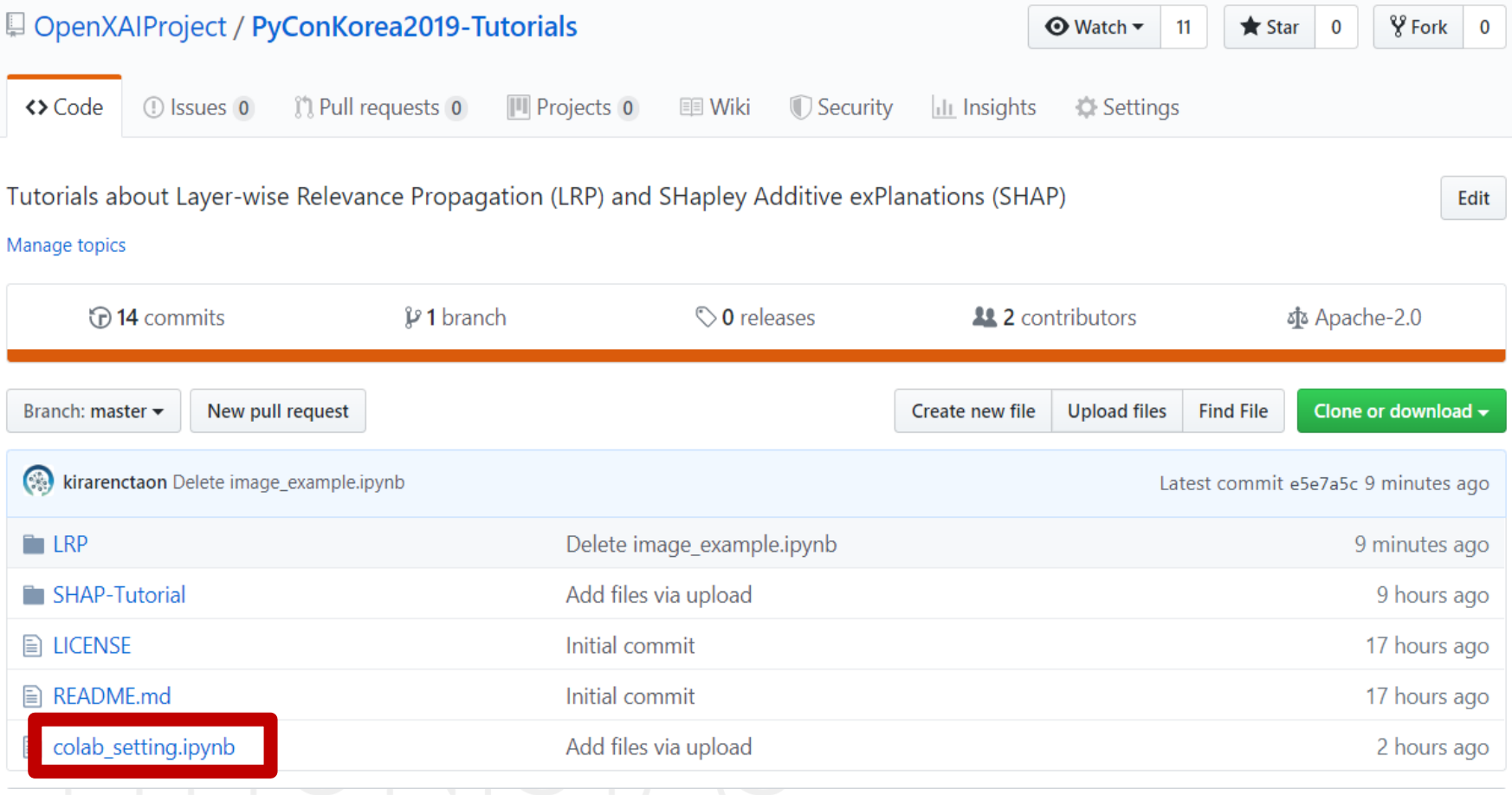
1. 튜토리얼 안내 페이지 접속

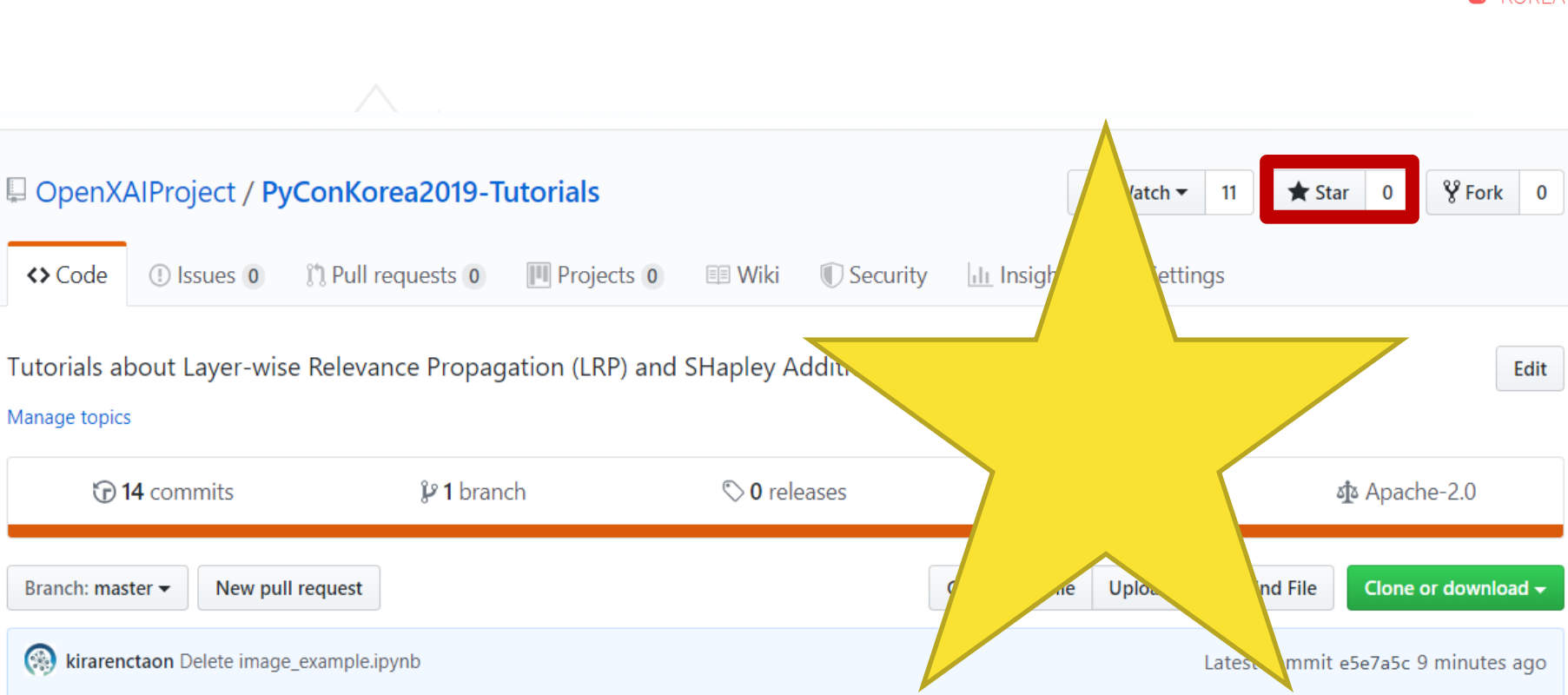
http://xai.unist.ac.kr/ >event>2019 PYCON TUTORIAL

2. 구글에 로그인

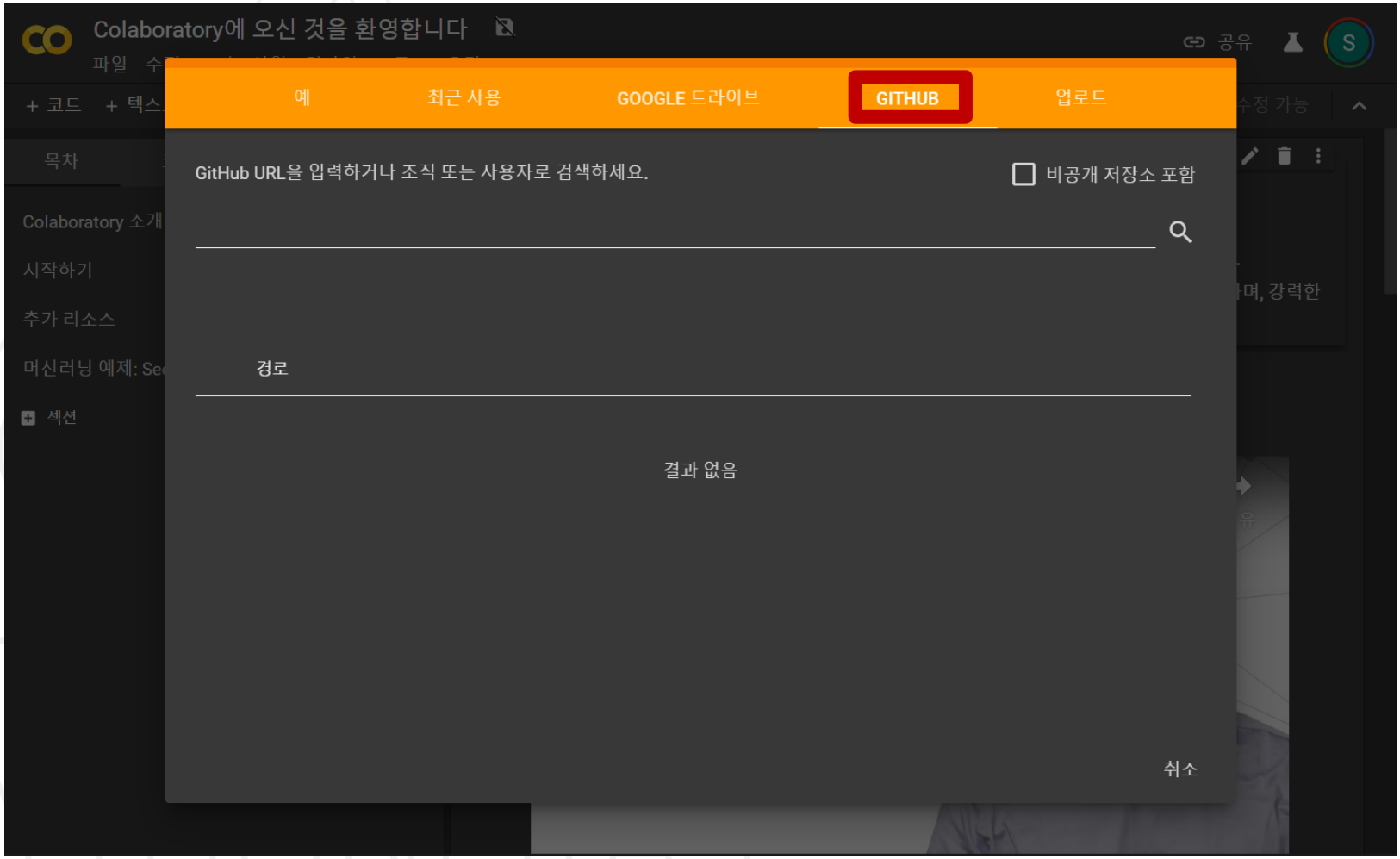
colab_setting.ipynb 파일의 주소를 복사하여,

https://github.com/OpenXAIProject/PyConKorea2019-Tutorials/blob/master/colab_setting.ipynb

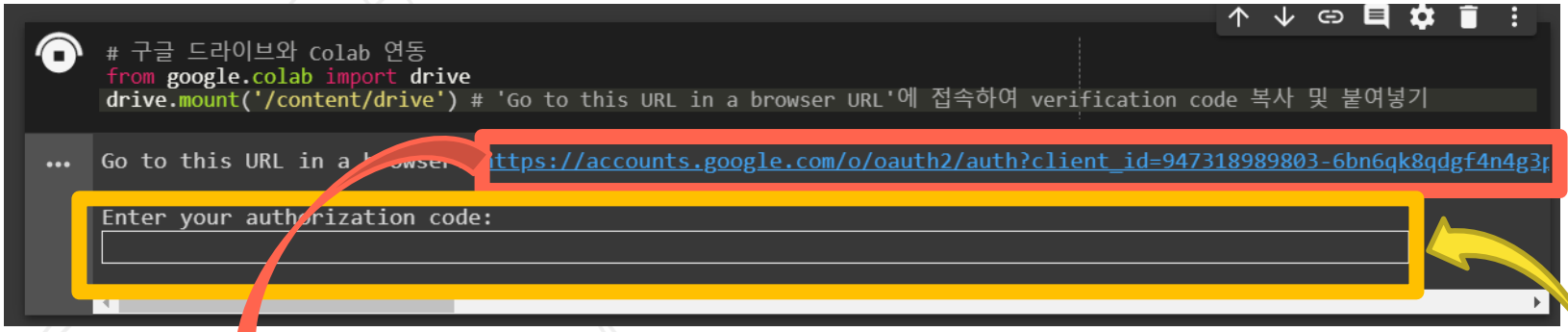


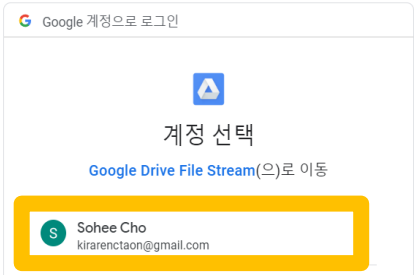

OpenXAIProject / PyConKorea2019-Tutorials
Watch 11
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Security
Insigh
ettings
Tutorials about Layer-wise Relevance Propagation (LRP) and SHapley Addit
Edit
Manage topics
14 commits
1 branch
0 releases
Apache-2.0
Branch: master
New pull request
Upload
nd File
Clone or download
kirarenctaon Delete image_example.ipynb
Latest mmit e5e7a5c 9 minutes ago
LRP
Delete image_example.ipynb
9 minutes ago
SHAP-Tutorial
Add files via upload
9 hours ago
LICENSE
Initial commit
17 hours ago
README.md
Initial commit
17 hours ago
colab_setting.ipynb
Add files via upload
2 hours ago

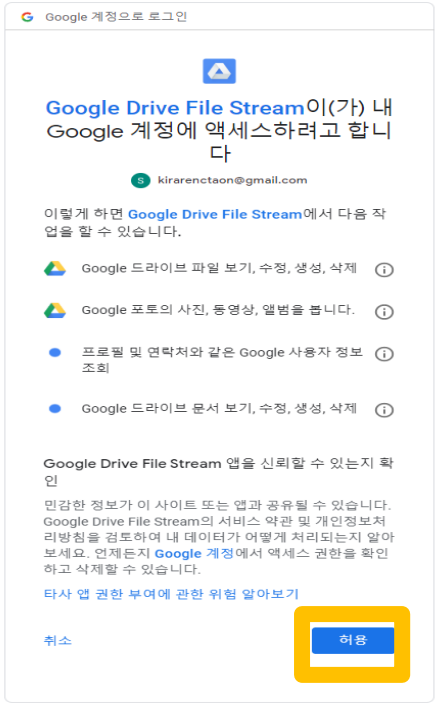
구글 코랩 GITHUB 탭에 복사해 온 주소를 입력합니다.

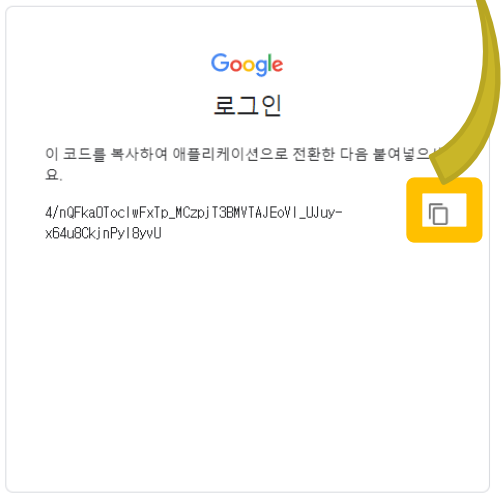


구글 드라이브에 마운팅하기 위한 인증 실행해주세요.









만약 아래와 같은 안내가 나온다면 ...

```
[116] # 구글 드라이브와 Colab 연동
      from google.colab import drive
      drive.mount('/content/drive') # 'Go to this URL in a browser URL'에 접속하여 이용동의 후 인증코드 복사해서 붙여넣기

      Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

이미 연동된 것으로 그대로 진행하면 됩니다.

연동 후에는 아래 코드를 모두 실행해
오늘 사용할 코드를 구글 드라이브에 클론합니다.