고등과학원
초학제 연구프로그램
올해의 주제 2019년도 연구
인공지능과 포스트휴머니즘

제1회 워크숍

설명가능 인공지능의 발전과 그 사회적 함의

◆ 일시: 2019년 3월 29일(금) 15:00 ~ 17:40
◆ 장소: 고등과학원 8호관 1층 세미나실 (8101호)
◆ 주관: 인공지능과 포스트휴머니즘 연구단
◆ 주최: 고등과학원 초학제 연구프로그램



Kias 로고
고등과학원 초학제연구단
인공지능과 포스트휴머니즘

고등과학원
올해의 주제 2019년도 연구
초학제 연구프로그램
인공지능과 포스트휴머니즘

초 대 장

고등과학원에서 운영하는 **<초학제 연구프로그램>**의 '올해의 주제' 2019년도 연구를 진행하고 있는 **<인공지능과 포스트휴머니즘>** 연구단입니다. 오늘날 인공지능로봇 기술이 급속도로 발전하면서 사회관계 및 생활세계의 변화뿐 아니라, 그에 수반하는 윤리적인 법적인 쟁점들의 부상과 더불어, 휴머니즘 및 인간 정체성에서도 많은 변화가 예견되고 있습니다.

본 연구단은 이러한 시대적 조망 하에 인공지능로봇 기술을 중심으로 우선 그것의 발전 배경과 현황 및 전망을 구체적으로 살펴보고, 그것이 인류 사회에 던지는 사회적 함의를 탐색하고자 합니다. 나아가 인간-인공지능 간 새로운 방식의 상호작용과 인류 사회의 변화, 이에 수반하는 윤리적·법적·사회적 쟁점들, 그리고 인공지능의 존재적 본질과 인간 정체성에 대한 이해의 변화와 같은 인간학적 주제들도 함께 논의해 보고자 합니다. 이를 바탕으로 궁극적으로 인간과 인공지능이 함께 공존하면서 발전할 수 있도록 하는 적합한 담론으로서의 포스트-휴머니즘을 모색해 보고자 합니다.

이를 위해 인공지능 분야의 핵심기술을 중심으로 해당 분야의 발전 배경과 현황 및 전망, 그리고 사회적·인간학적 함의 등을 주제로, 관련 전문가들을 모시고 학제 간 학문적 교류 및 소통을 위한 학술 워크숍을 개최하려 합니다. 본 연구단의 학술 워크숍에 많은 관심과 참여를 바랍니다.

<인공지능과 포스트휴머니즘> 연구단
연구책임자 이중원



고등과학원
초학제 연구프로그램
올해의 주제 2019년도 연구
인공지능과 포스트휴머니즘

제1회 워크숍

설명가능 인공지능의 발전과 그 사회적 함의

◆ 일시: 2019년 3월 29일(금) 15:00 ~ 17:40
◆ 장소: 고등과학원 8호관 1층 세미나실 (8101호)
◆ 주관: 인공지능과 포스트휴머니즘 연구단
◆ 주최: 고등과학원 초학제 연구프로그램

시 간	발표자	제 목
15:00~15:10	이중원(서울시립대학교)	개 회 사
발 표		사회: 최인령(고등과학원)
15:10~15:50	최재식(울산과학기술원)	**설명가능 인공지능의 최신 연구동향과 활용 전망**
15:50~16:30	전진권(명지대학교)	**설명가능 인공지능 : 사회를 위한 도구가 될 것인가?**
질의응답 및 토론		사회: 이중원(서울시립대학교)
16:50~17:10	이상욱(한양대학교)	종합논평
17:10~17:40	질의응답 및 토론	

최 재 식	전 진 권
울산과학기술원(UNIST) 설명가능 인공지능 연구센터장을 맡고 계시는 최재식 교수님께서 <설명가능 인공지능의 최신 연구동향과 활용 전망>이라는 제목으로 빅 데이터의 인과관계를 분석해서 인공지능이 내린 판단에 대해 합리적인 이유와 근거를 제시하여 사람이 이해할 수 있도록 설명할 수 있는 인공지능의 개발 현황과 전망에 대해 발표해주시겠습니다.	명지대학교 과학철학자 전진권 교수님께서 <설명가능 인공지능: 사회를 위한 도구가 될 것인가?>라는 제목으로 설명가능 인공지능에 대한 인간학적 관점과 사회적 함의에 대해 발표해 주시겠습니다. 전진권 교수님은 인지과학, 진화론, 인공지능 등의 첨단 과학 분야와 인문학이 만나는 접점에 대해 연구하고 설명가능 인공지능에 대해서도 철학적으로 접근하고 있습니다.



Innovative Growth Engine Project

Ministry of Science and ICT

설명가능인공지능 연구동향

UNIST 설명가능인공지능 연구센터

최재식

2019. 03. 29

http://www.openXAI.org/

■ 주관연구기관: 울산과학기술원

■ 참여연구기관: KAIST 한국과학기술원, 고려대학교, 연세대학교 의료원 YONSEI UNIVERSITY HEALTH SYSTEM, AITRICS

■ 해외협력기관: Microsoft Research, Penn, Michigan State University, UC Davis, Johns Hopkins University, The University of North Carolina at Chapel Hill, 中國科學院 Chinese Academy of Sciences

■ 미국 XAI 프로젝트 참여 협력기관: DARPA, MIT Massachusetts Institute of Technology, UCLA, Carnegie Mellon University, SRI International

■ 기술이전의향기업: NAVER, POSCO, BNK 경남은행, AI ZEN

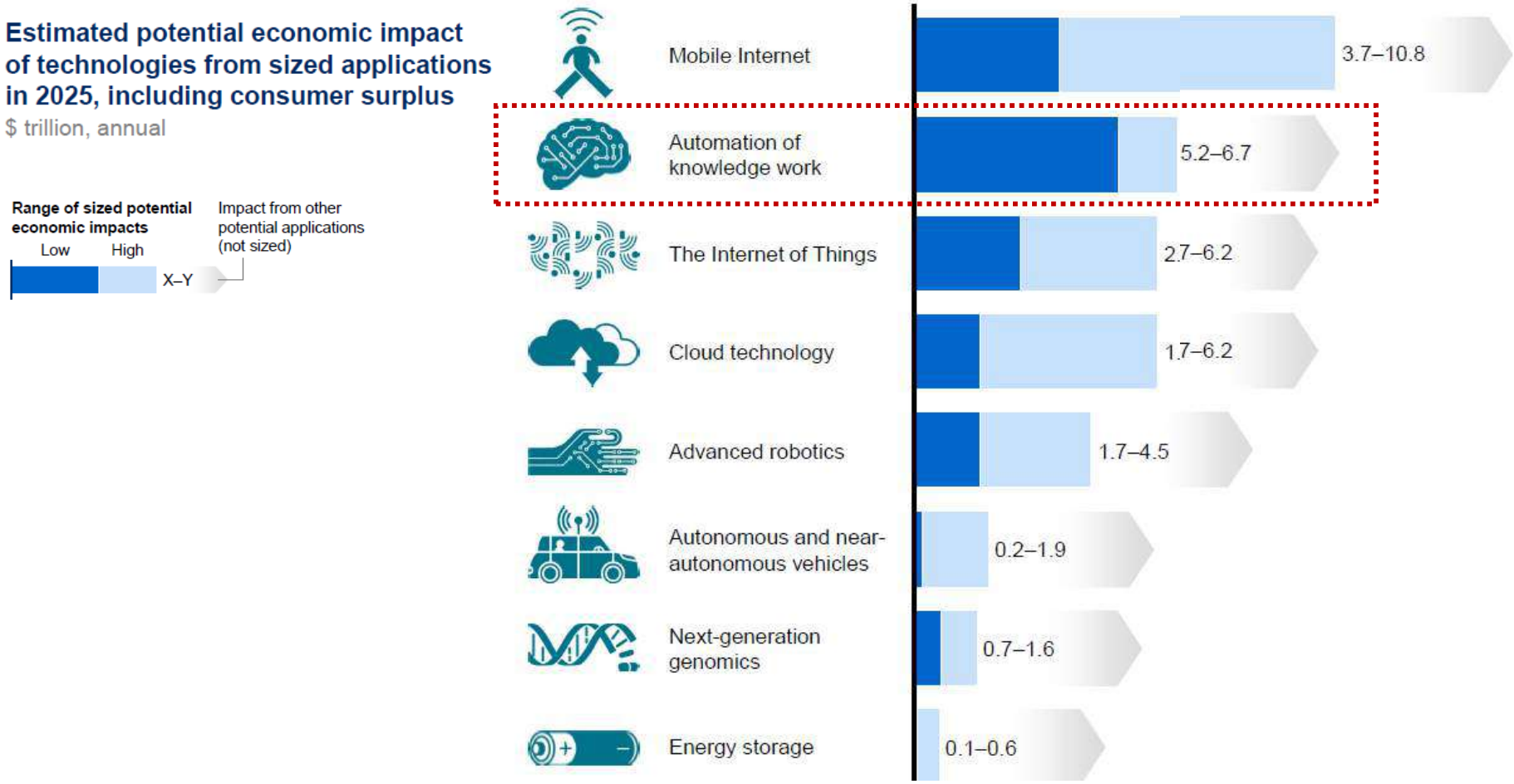
목차

소개

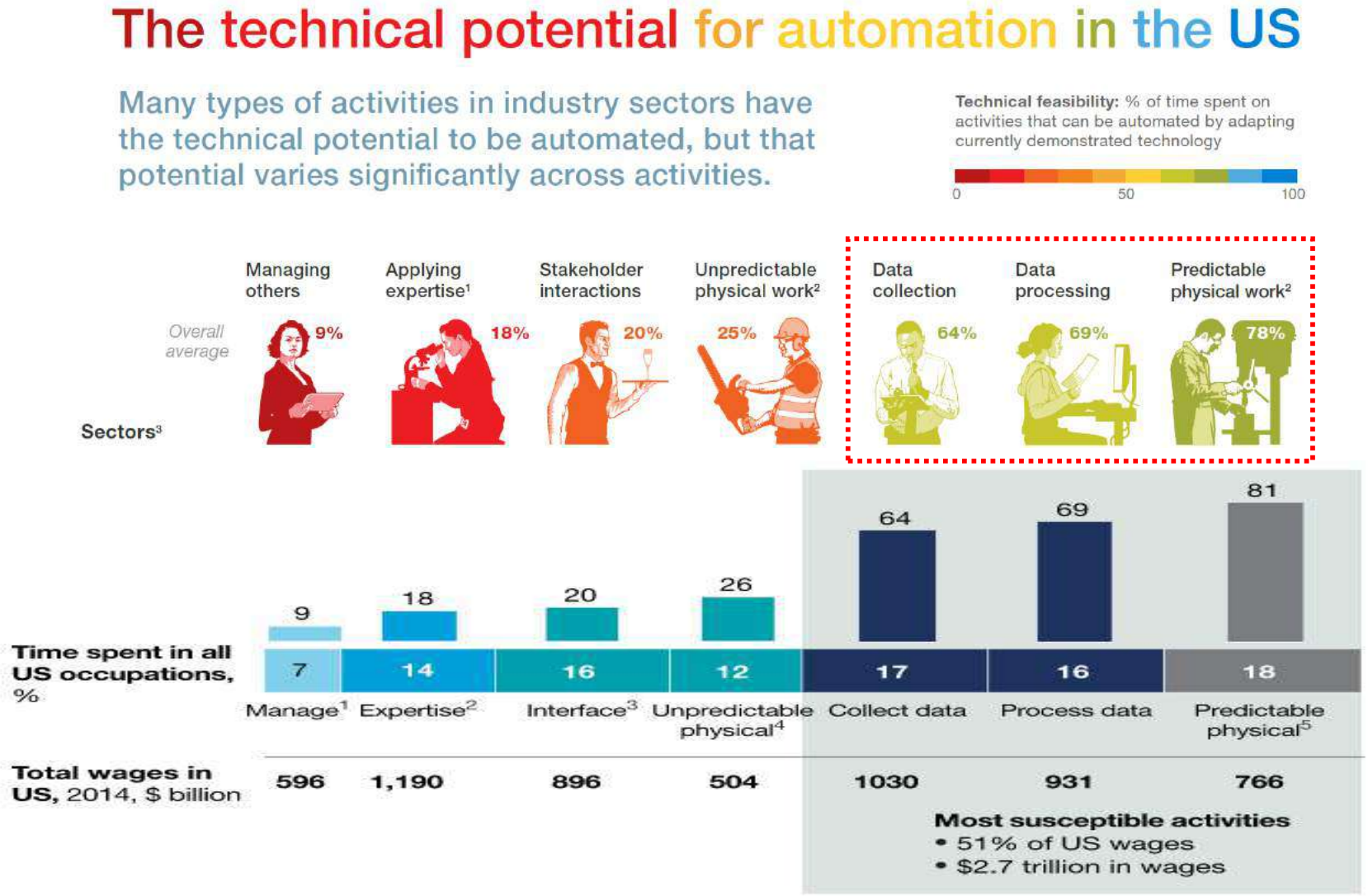
미국의 DARPA의 XAI 프로젝트

과학기술정보통신부 설명가능인공지능 연구센터

향후 계획



Estimated Economic Impact of Disruptive Technology [McKinsey 2013]



Automation of Knowledge Work [McKinsey 2013]



의료

IBM Watson의 위암 진단 정확도 (미국: 93% -> 한국: 49%)

- 미국 Memorial Sloan Kettering의 데이터가 한국과 다름

MD Anderson은 IBM Watson기반 암진단 프로젝트 중단(2017년 2월)

- $62M 투자 후 4년만에 중단. 데이터의 부족이 주요 원인



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

시사점

의료 분야와 같이 데이터가 부족한 경우 효과적으로 인과관계를 학습하는 원샷/제로샷 기술이 필요함

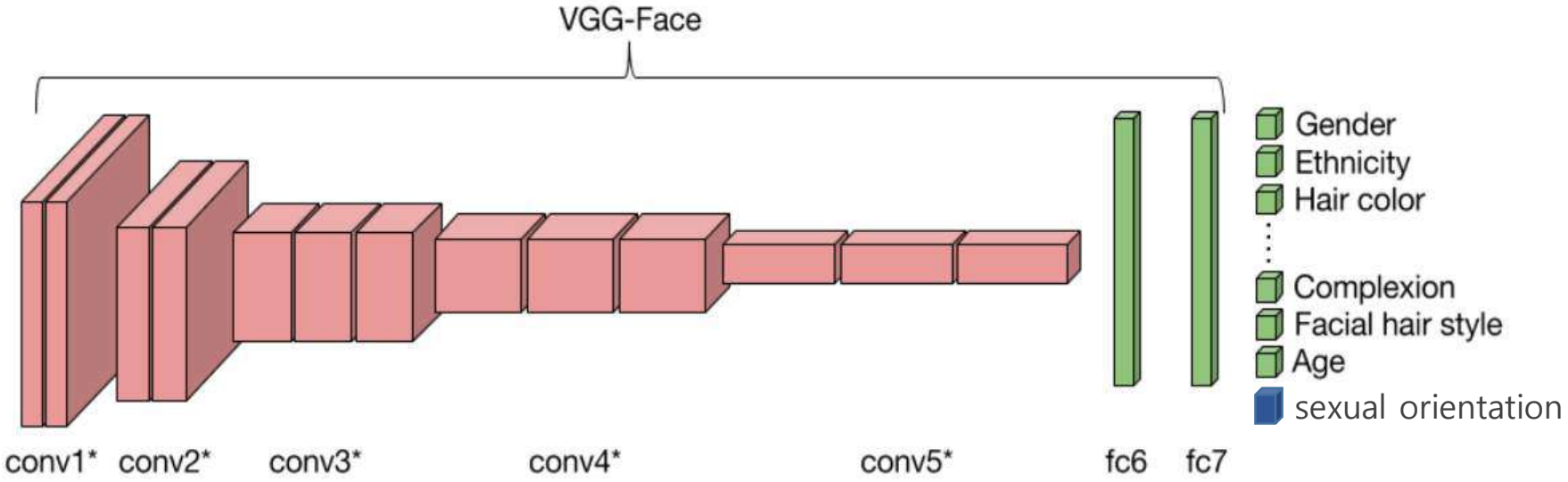
인공지능 기술을 주요 의사결정에 적용하기 위해서 법적으로 이유 제시 기능이 필요함

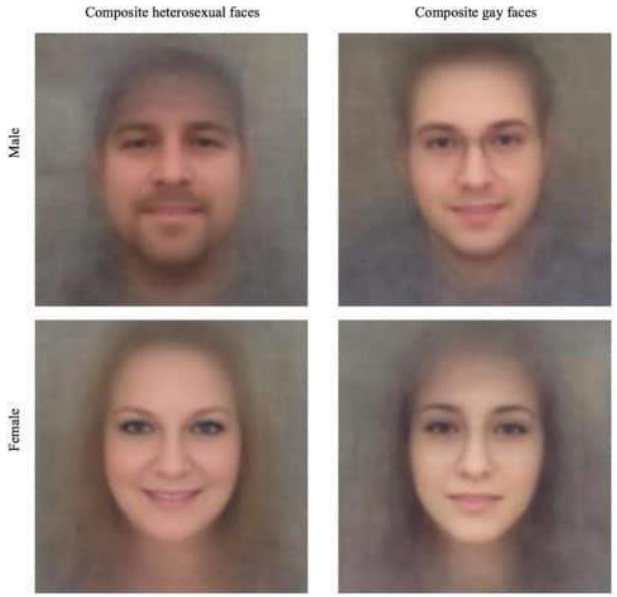
https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing

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%

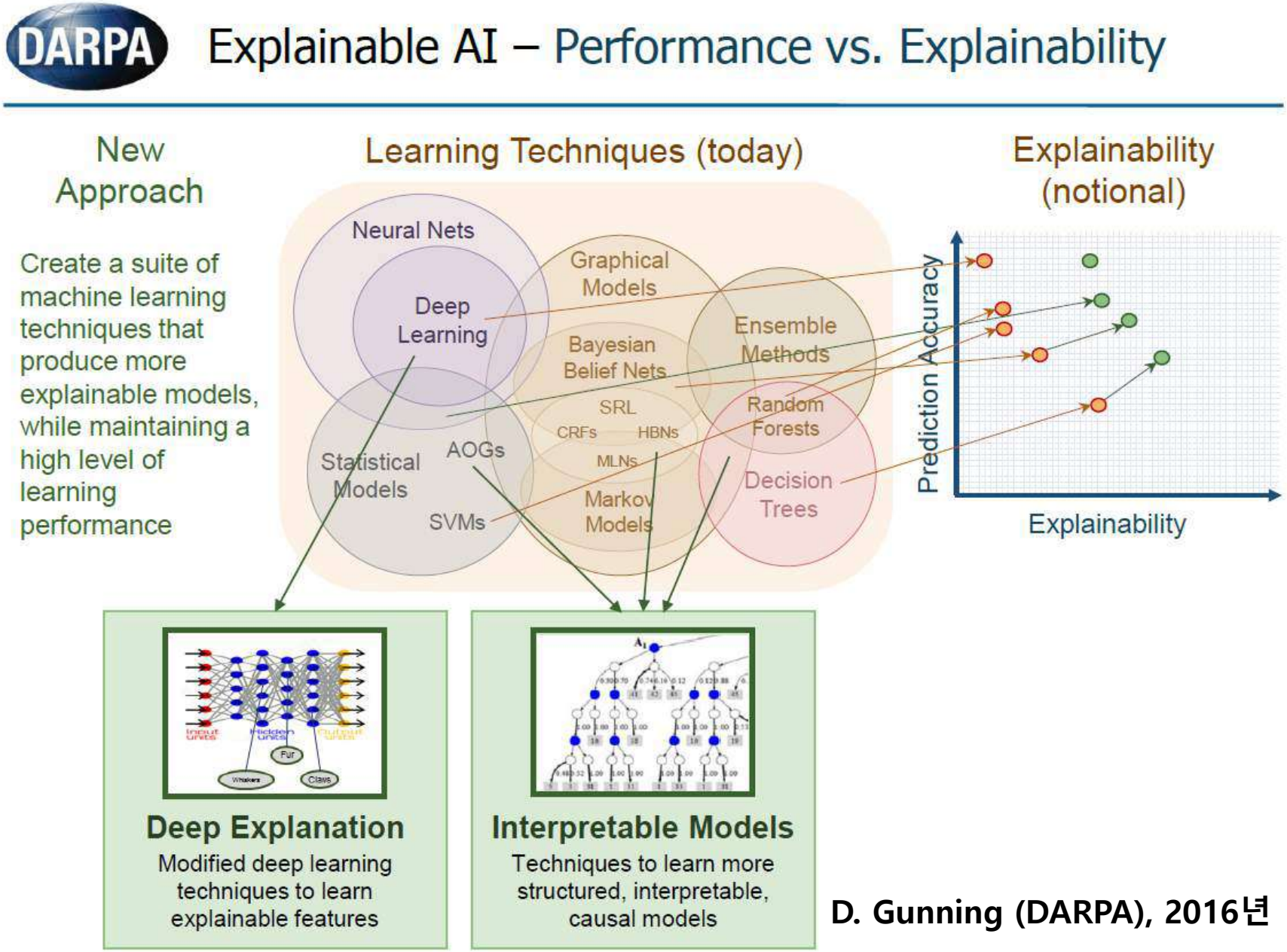
VS

VGG-Face의 정확도: 83 ~ 91%

문제: VGG-Face는 결정의 이유를 설명할 수 없음!

M. Kosinski 교수 연구팀 (Stanford), 2017년

해결방법 - DARPA 설명가능 인공지능



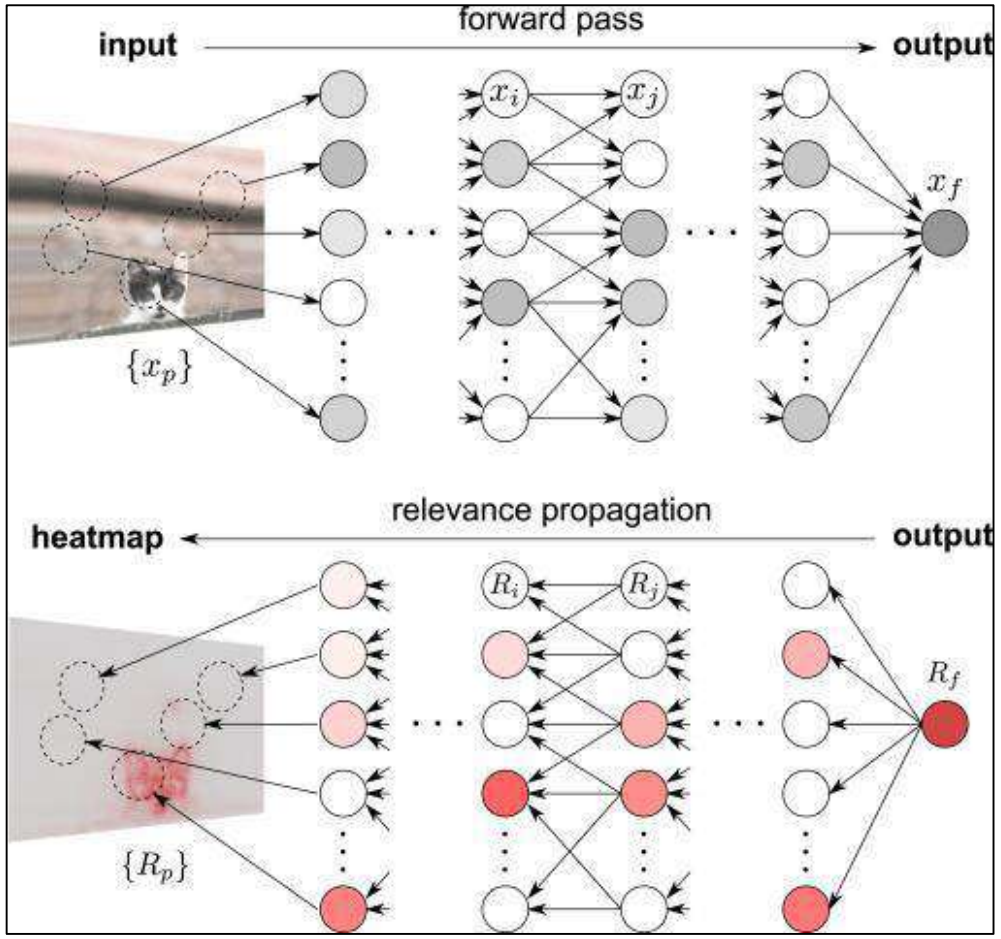
D. Gunning (DARPA), 2016년

응용사례 - 설명가능 딥러닝 연구

Layer-wise Relevance Propagation(LRP)



딥러닝 모델의 예측 결과에 대해 시각적 해석을 돕는 인터페이스 Heatmapping을 개발하여 공개 (www.heatmapping.org)

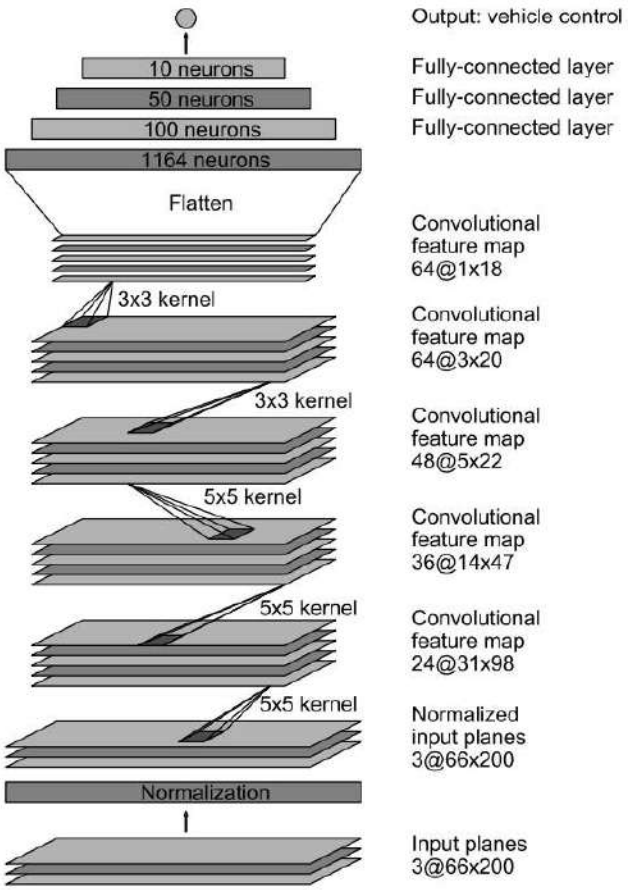


딥러닝 예측 결과 이유 설명을 위한 LRP 알고리즘 [CVPR, 2015]

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

응용사례 - 설명가능 딥러닝 응용

NVIDIA의 PilotNet(자율주행 딥러닝)의 결정을 설명하는 딥러닝



PilotNet 구조도



자율주행의 이유를 설명해주는 딥러닝 기술

NVIDIA/Google 연구팀, 2017년

설명가능 인공지능에 관한 수요

설명가능 인공지능 수요 그룹에 따른 분류

- 인공지능 개발자
  - **응용 프로그램을 개발 주체**
  - 개발중인 인공지능 모델의 디버깅을 목표
  - Traceability - did the thing right
  - 설명성(explainability)/해석성(interpretability) 사용

  예, 상대적 속성 전파(Relative Attributing Propagation)
  희소 특징 추출(Sparse Feature Extraction)

- 인공지능 이론가
  - **고급 AI 이론 연구 주체**
  - 인공지능 이론 분석 및 복잡 공간의 해석이 목표
  - 해석성(interpretability) 사용

  예, 커널 분할(Kernel decomposition),
  특징의 불확실성 계산(Feature uncertainty analysis)

- 인공지능 윤리/법률
  - **인공지능의 공정성 및 투명성 확보가 목표**
  - Assurance - does the right thing
  - 알기쉬움(intelligibility), 읽기쉬움(legibility) 사용

  예, EU의 일반 개인정보보호법, 의료 정보 DB

- 인공지능 사용자
  - **인공지능 시스템 사용자**
  - 인공지능 산출물에 영향을 받는 사람
  - Justification - did the right thing

  예, 의료, 금융 자동 보고서 시스템, 인과관계 추출

[A. Preece et al., "Stakeholders in Explainable AI", Arxiv, 2018]

목차

소개

미국의 DARPA의 XAI 프로젝트

과학기술정보통신부 설명가능인공지능 연구센터

향후 계획

Explainable Artificial Intelligence (XAI)



Principle Investigators Meeting
12-14 FEB 2019

Program Manager: David Gunning
Information Innovation Office (I2O)
Defense Advanced Research Projects Agency (DARPA)

Approved for public release: distribution unlimited.


ECONOMY
NATIONAL SECURITY
BUDGET
IMMIGRATION
THE OPIOID CRISIS
EXECUTIVE ORDERS
Executive Order on Maintaining American Leadership in Artificial Intelligence
INFRASTRUCTURE & TECHNOLOGY | Issued on: February 11, 2019
SHARE:
ALL NEWS
By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:
Section 1. Policy and Principles. Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national

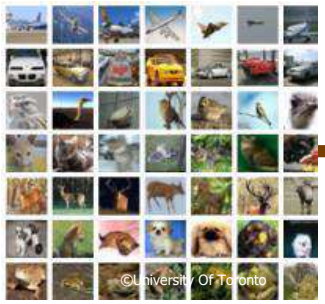
How is it done today?

Today



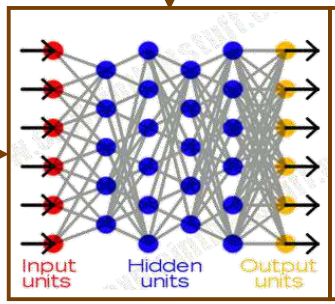
Input

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?



Training Data

Learning Process



Learned Function

This is a cat
(p = .93)

Output



User

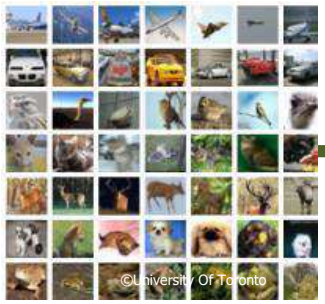
What are we trying to do?

Tomorrow



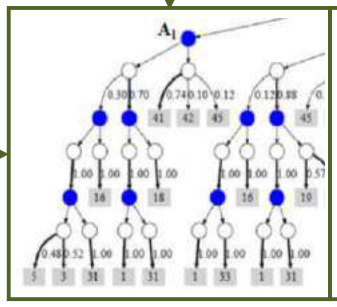
Input

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred



Training Data

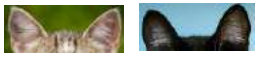
New Learning Process



Explainable Model

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

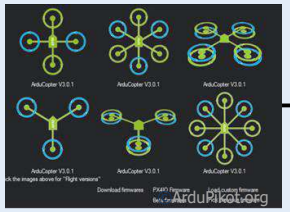


Explanation Interface



User

Challenge Problems

	Learn a model	Explain decisions	Perform a task	
Data Analytics Classification Learning Task	 Multimedia Data Classifies items of interest in large data set	Explainable Model → Explanation Interface Explains why/why not for recommended items	Recommend Explanation  Analyst decides which items to report, pursue	An analyst is searching for items of interest in massive multimedia data sets
Autonomy Sequential Learning Task	 ArduPilot & SITL Simulation Learns decision policies for missions	Explainable Model → Explanation Interface Explains behavior in an after-action review	Actions Explanation  Operator decides which future tasks to delegate	An operator is directing autonomous systems to perform a series of missions

Approved for public release: distribution unlimited.

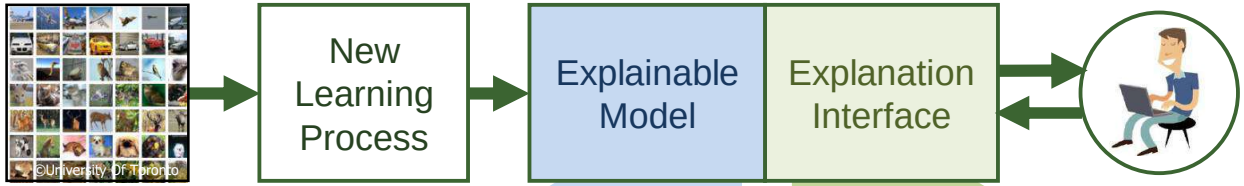
Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems

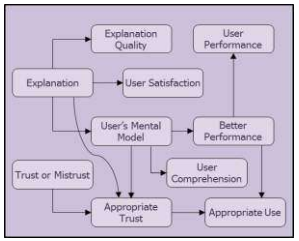


Approved for public release: distribution unlimited.

XAI Performers & Technical Approaches



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



IHMC Psychological Models of Explanation

Randomized Input Sampling for Explanation

Neural Network Prediction

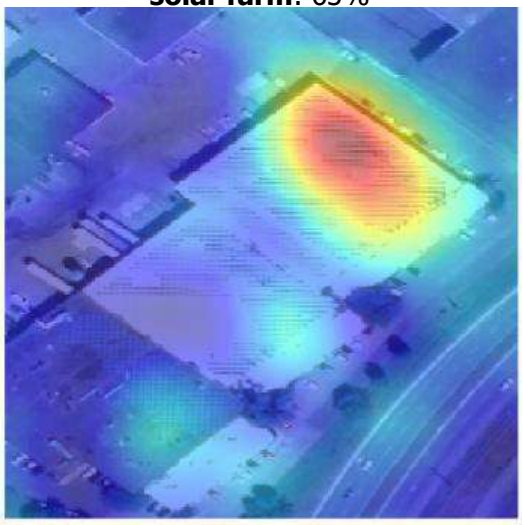
solar farm: 63%, **shopping mall**: 23%



Image from the FMoW dataset

RISE Explanation for **solar farm**

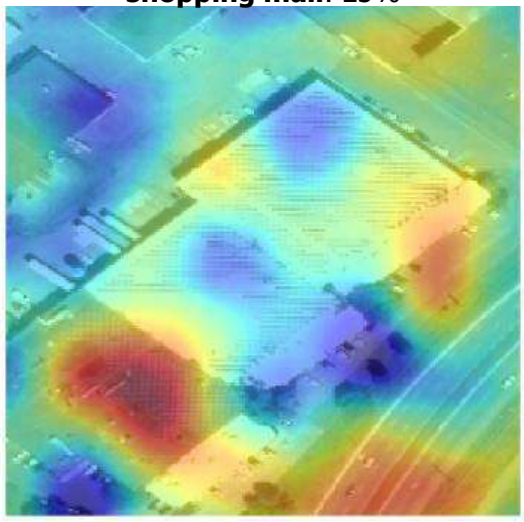
solar farm: 63%



Increasing importance

RISE Explanation for **shopping mall**

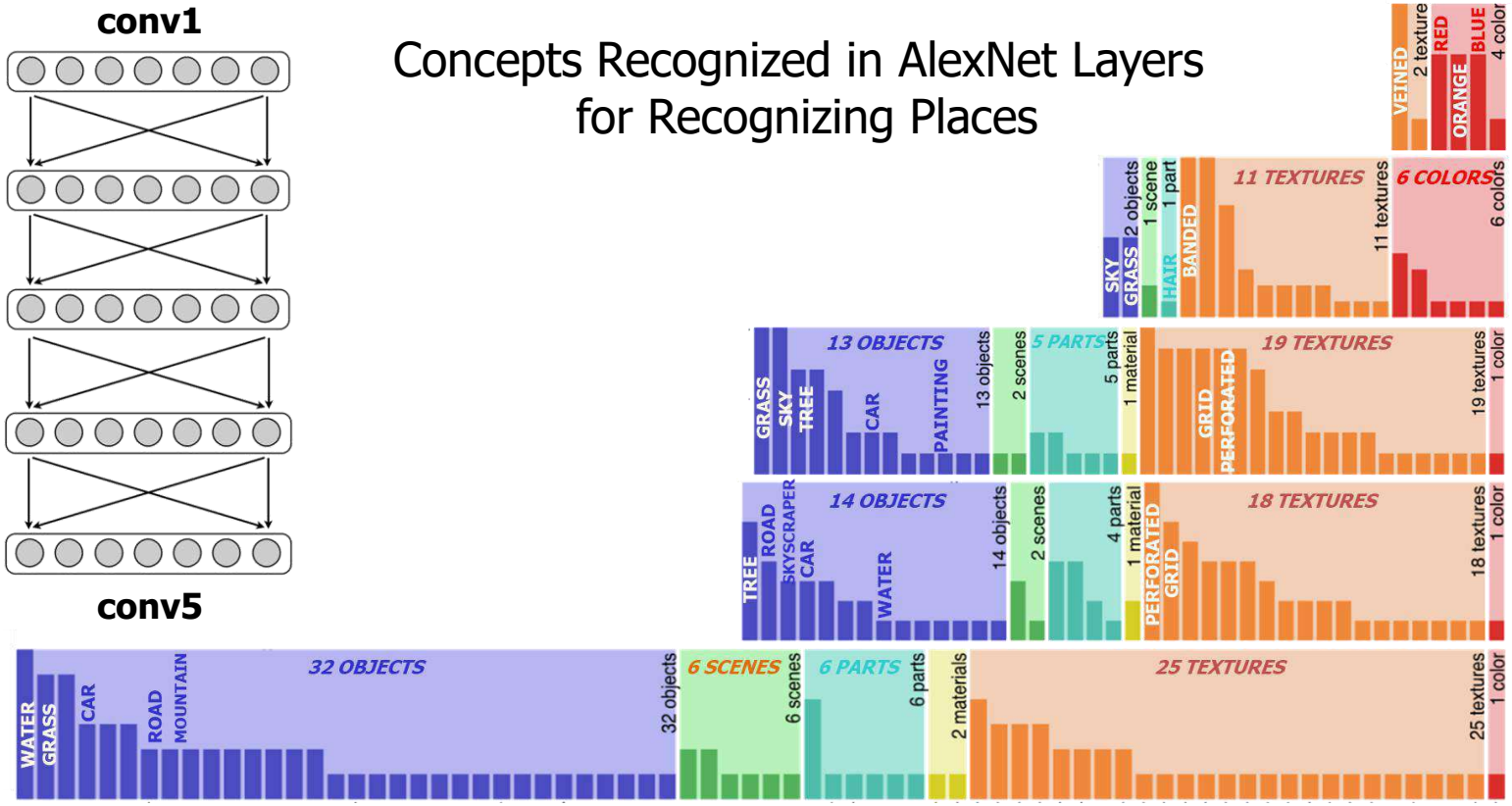
shopping mall: 23%



Boston University

Vitali Petsiuk, Abir Das, Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. Proceedings of the British Machine Vision Conference (BMVC), 2018.

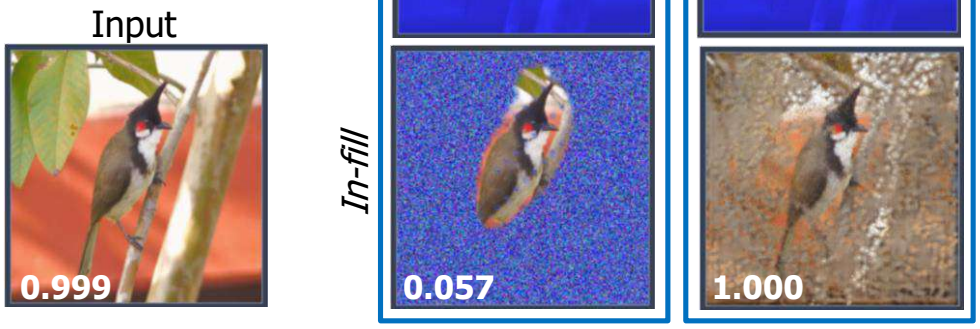
BBN/MIT – Network Dissection



SRI - DARE: Deep Attentional Representations for Explanation

Explaining Image Classifiers by Counterfactual Generation



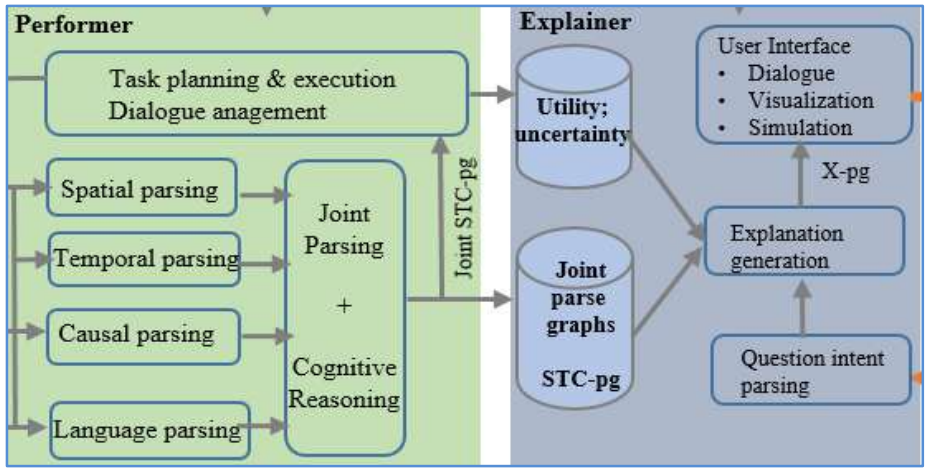


University of Toronto, Vector Institute

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining Image Classifiers by Counterfactual Generation (NeurIPS 2018)

UCLA, Oregon State, Michigan State

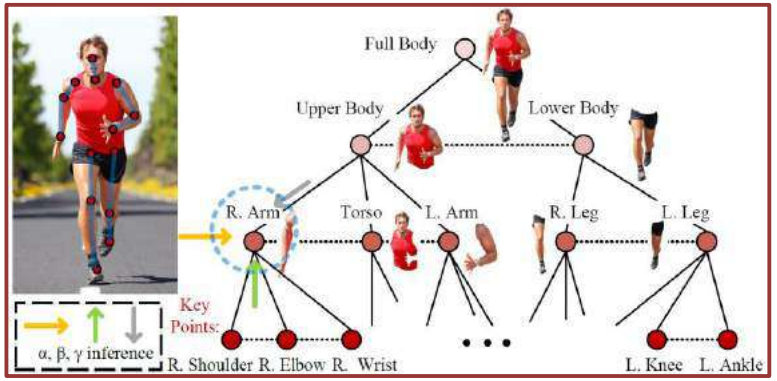
System Architecture



Performer
Outputs interpretable representations in a spatial, temporal, and causal parse graph (STC-PG)

Explainer
Outputs an explanatory parse graph (X-PG) in a dialogue process

STC Parse Graph



An attributed parse graph for a running person. Each node has 3 computing channels:

- α: grounding the node on DNN features;
- β: bottom-up;
- γ: top-down.

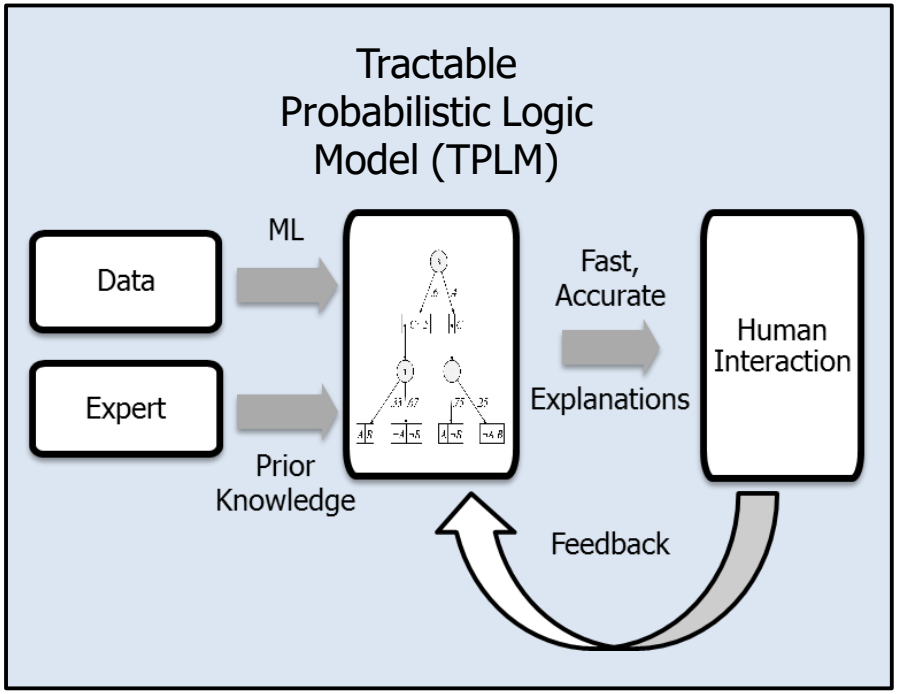
An explanation is represented as parse graph X-pg

Approved for public release: distribution unlimited.

Tractable Probabilistic Logic Models

UT Dallas, UCLA, Texas A&M, Indian Institute of Technology

Use interpretable and tractable models based on well-founded principles from logic and probability theory



Find all videos in which a person peels oranges
(explanations are captions generated by TPLMs)



Person using his hands to peel oranges. I can see the orange skin



Person using his hands. I can see the orange skin and skinless orange



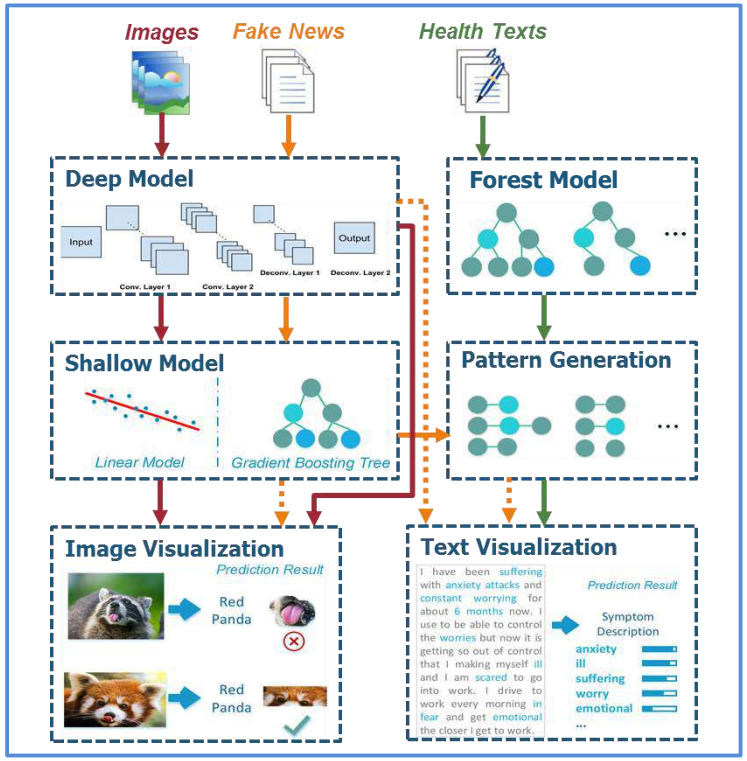
Person using his hands to peel oranges. I can see the orange skin on the table and peeled oranges

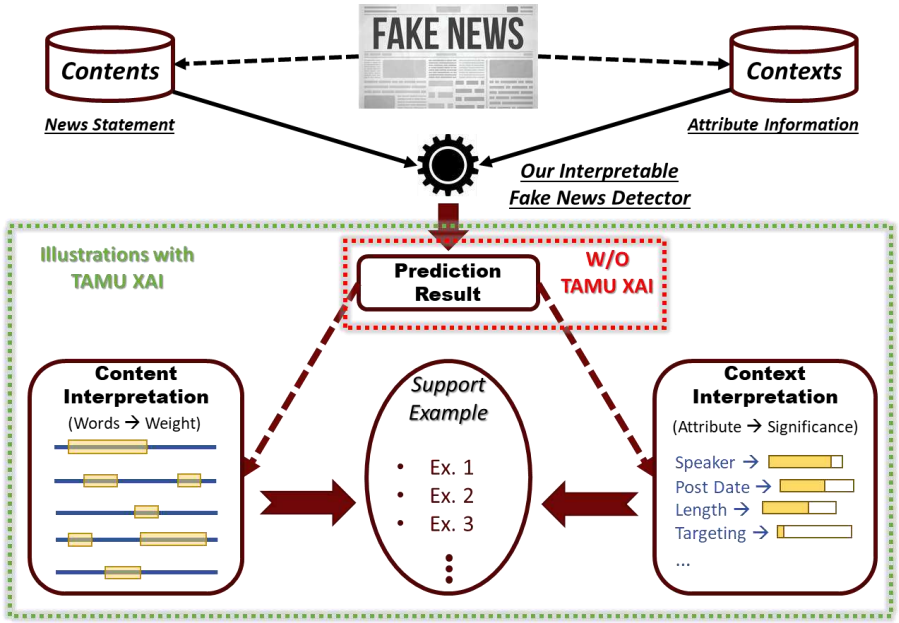
Approved for public release: distribution unlimited.

Transforming Deep Learning to Harness the Interpretability of Shallow Models

Texas A&M, Washington State

Develop an end-to-end interpretable deep learning infrastructure with image and text datasets



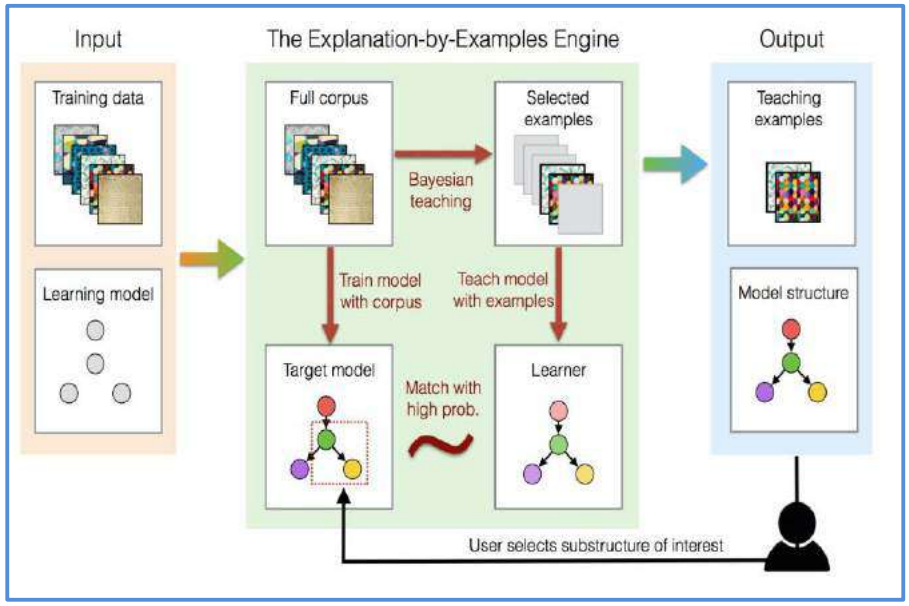


Approved for public release: distribution unlimited.

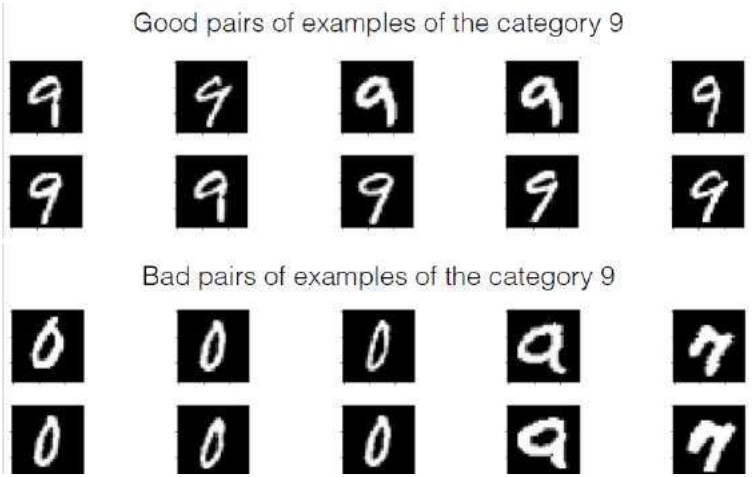
Model Explanation by Optimal Selection of Examples

Rutgers University

Extend Bayesian teaching to enable automatic explanation by selecting the subset of data that are most representative of the model's generative process



Good and bad examples for teaching a category (illustrates model strengths and weaknesses)



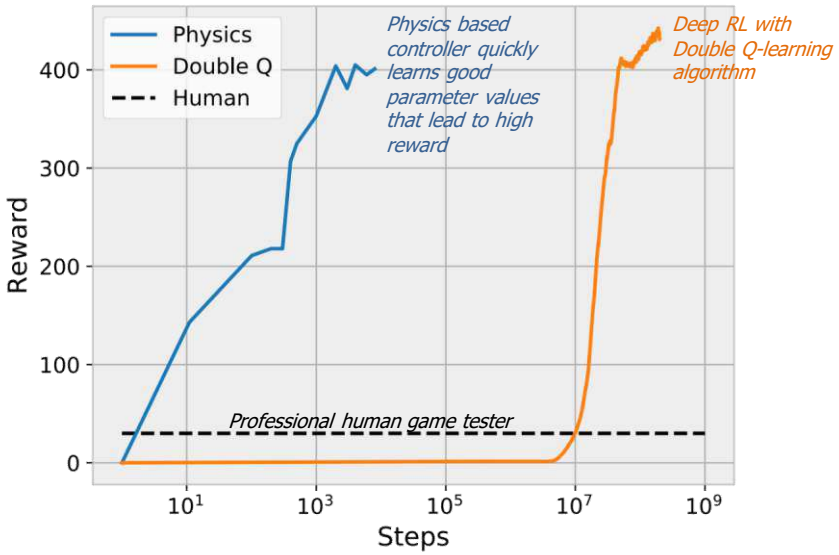
Approved for public release: distribution unlimited.

CMU - XRL: Explainable Reinforcement Learning for AI Autonomy

Differentiable Physics and Control



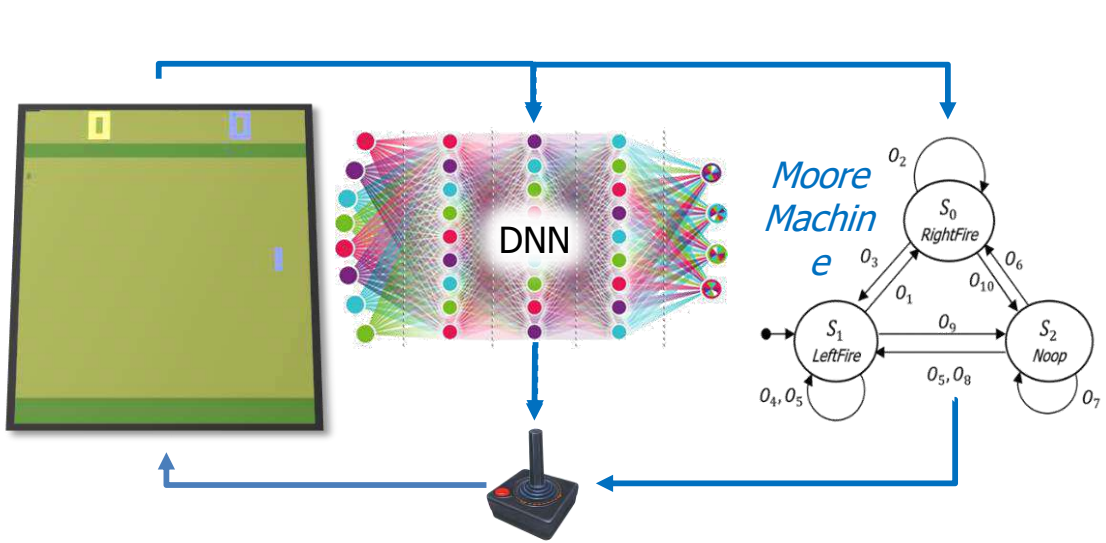
Carnegie Mellon University, Massachusetts Institute of Technology, and Bosch Center for AI

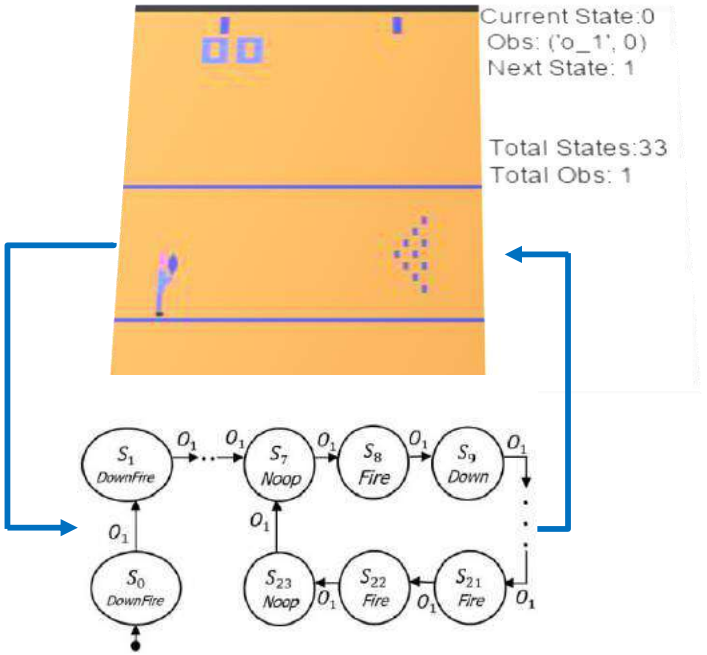


Filipe de A. Belbute-Peres, Kevin A. Smith, Kelsey R. Allen, Joshua B. Tenenbaum, and J. Zico Kolter. End-to-End Differentiable Physics for Learning and Control. NeurIPS 2018: 7178-7189.

OSU - xACT

Explaining RNN Memory Use

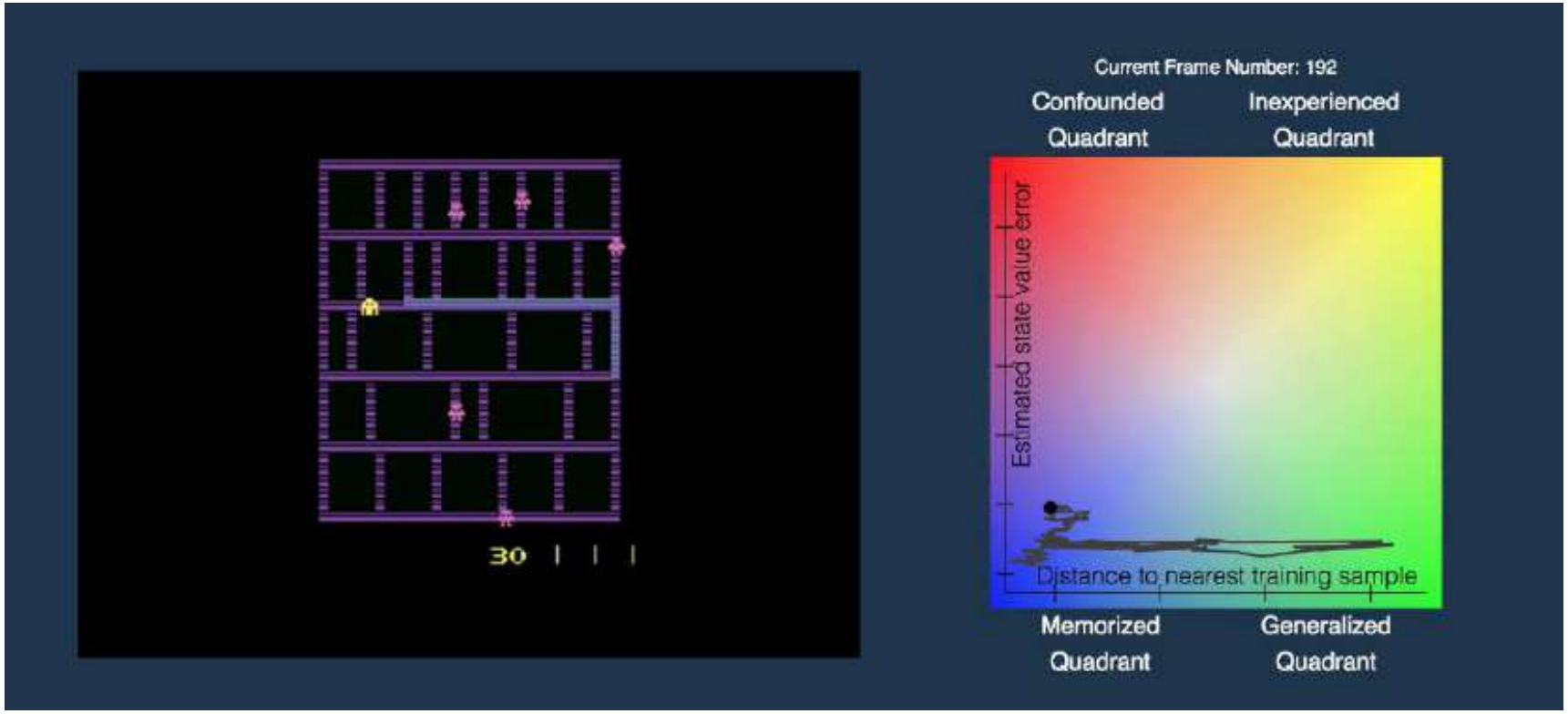




Oregon State University

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

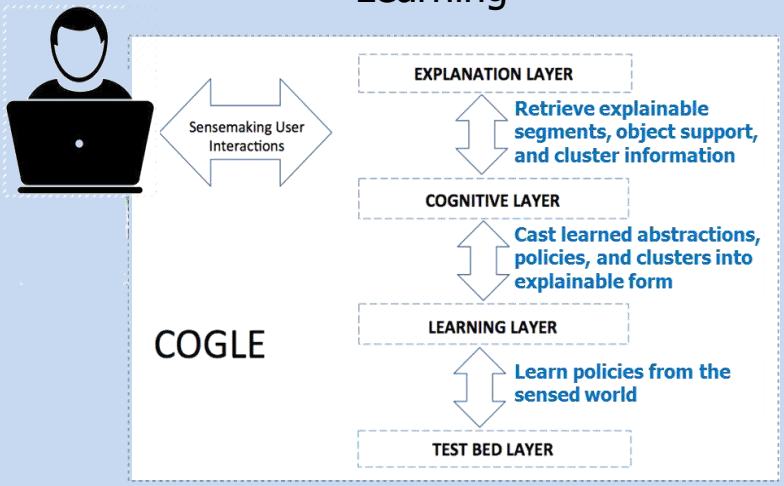
CRA - Causal Models to Explain Learning (CAMEL)



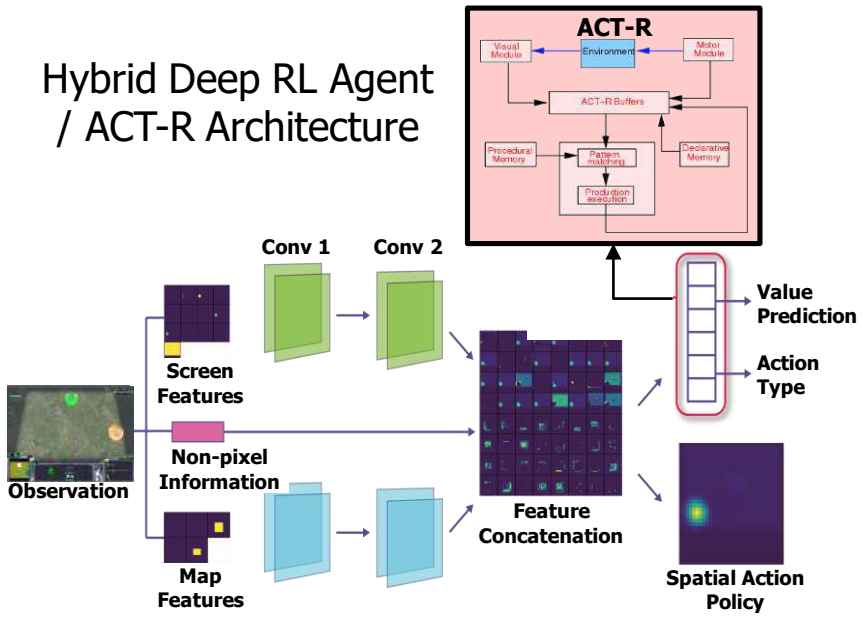
COGLE: Common Ground Learning and Explanation

PARC, CMU, U. Edinburgh, U. Michigan, USMA, IHMC

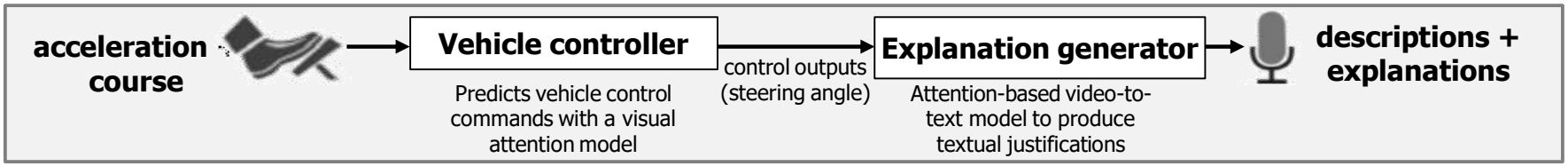
Layered Cognitive Architecture to Partition Explanation And Learning

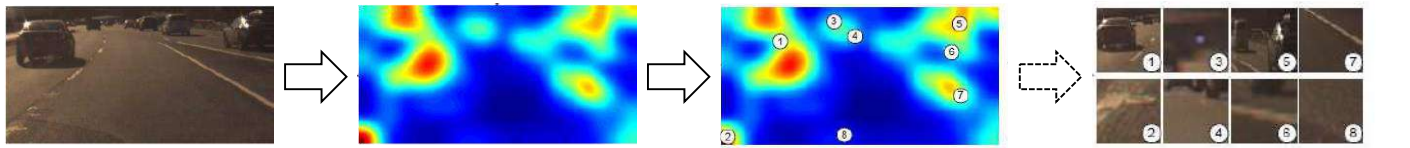


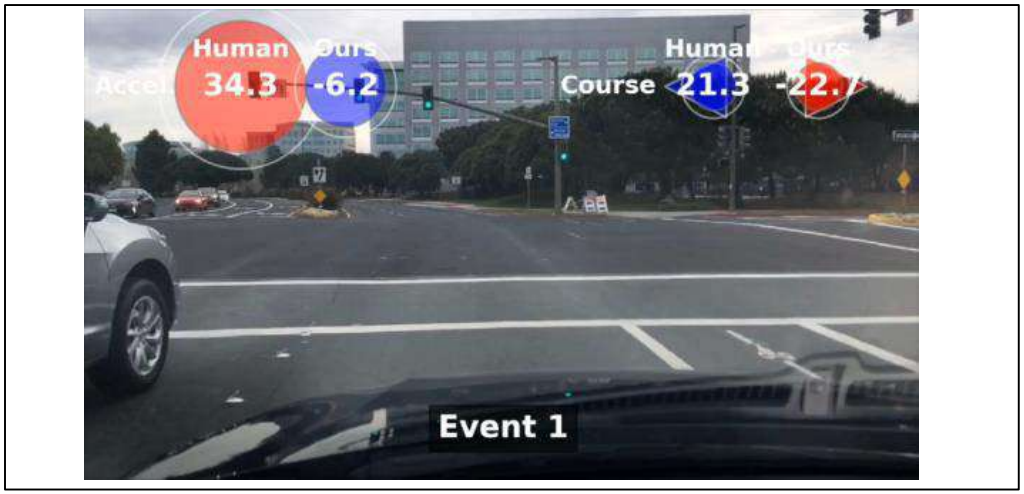
Hybrid Deep RL Agent / ACT-R Architecture



UCB - Causal Grounded Driving





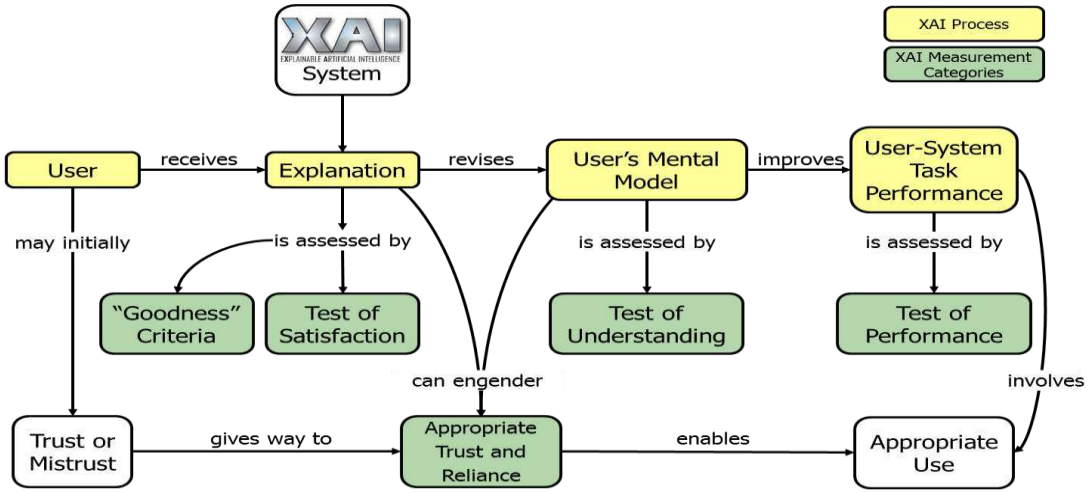


Kim and Canny, in ICCV, 2017
Kim, Rohrbach, Darrell, Canny, and Akata, in NIPS Interpretable ML Symposium, 2017

Approved for public release: distribution unlimited.

Experimental Framework

Explanation Process & Measures



Hoffman, Mueller, Klein, & Litman (2018). Metrics for Explainable AI: Challenges and Prospects. arXiv

Experimental Conditions

Without Explanation - The explainable learning system is used to perform a task without providing an explanation to the user

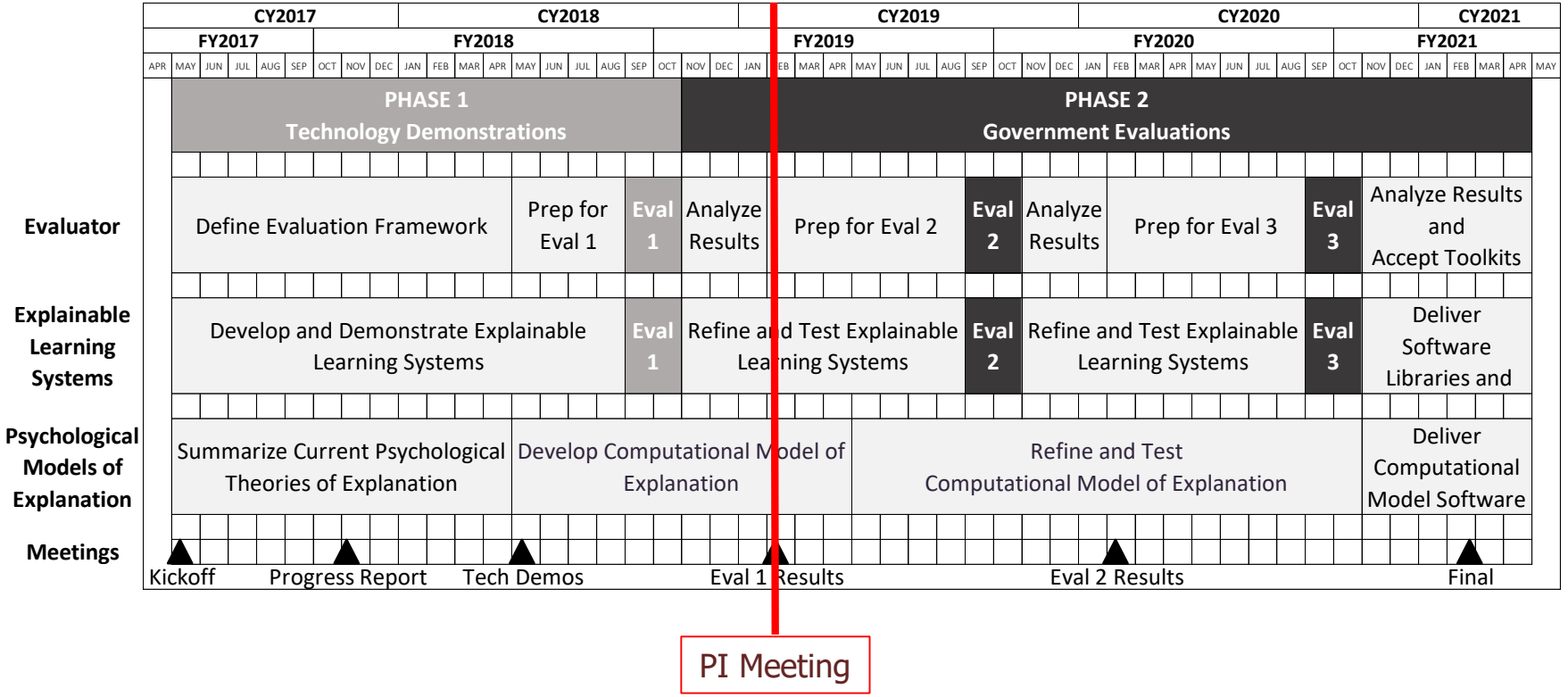
With Explanation - The explainable learning system is used to perform a task and generates explanations for every recommendation or decision it makes, and every action it takes

Partial Explanation - The explainable learning system is used to perform a task and generates only partial or ablated explanations (to assess various explanation features)

Control - A baseline state-of-the-art non-explainable system is used to perform a task

Approved for Public Release, Distribution Unlimited

XAI Program Schedule



Approved for public release: distribution unlimited.

목차

소개

미국의 DARPA의 XAI 프로젝트

과학기술정보통신부 설명가능인공지능 연구센터

향후 계획

Explainable AI Program in Korea

Goal Human-level Learning and Inference to overcome the limitations of Deep Neural Networks

AS-IS (Deep Learning)	TO-BE (Human-level Learning/Inference)
• It is hard to know the decision, so called **Blackbox** model • It does not work well when we **do not have enough training data**	• **Explainable learners** which can provide the reasons of decisions • Learning explainable models even with **data deficient environment**

Fund Institute of Information & Communication Technology Promotion (IITP) under Ministry of Science and ICT (MSICT)

as part of Innovative Growth Engine Project

Period July 2017 ~ December 2021 (54 months)

XAI Program Schedule in Korea

Phase 1

Competitive Demonstrations

Period: 2017.07 ~ 2018.12 (18 months)

Teams: 2

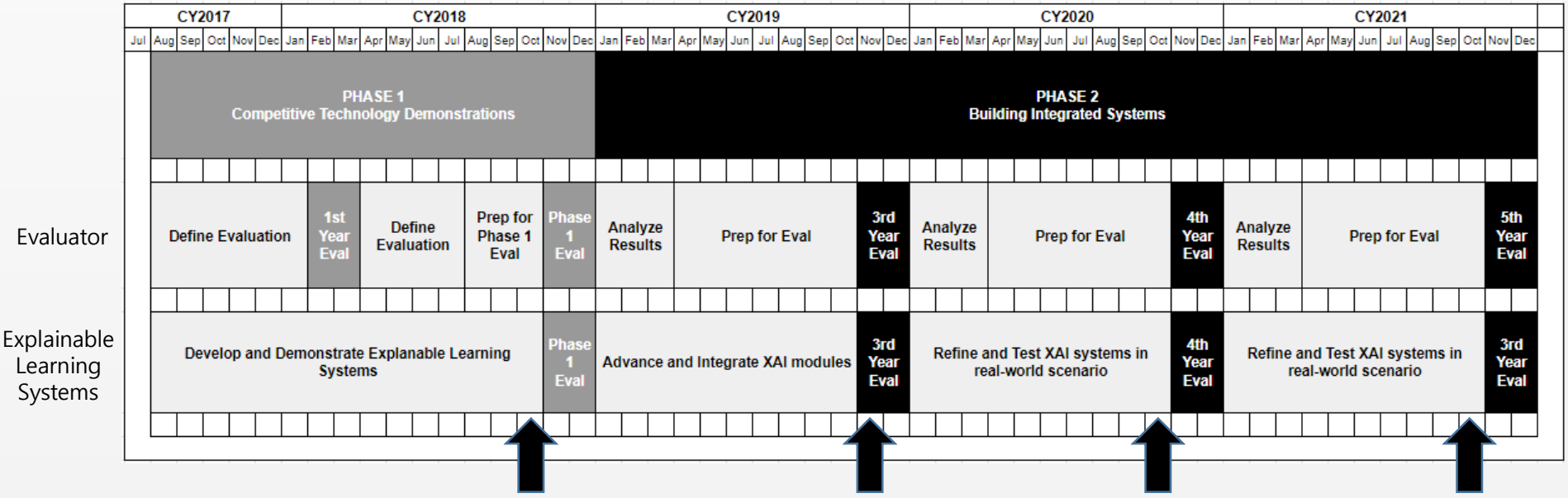
Phase 2

Building Integrated XAI Systems

Period: 2019.01 ~ 2021.12 (36 months)

Team: 1

Schedule



Project Organization

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
International Finance Trading / Medical Big Data Competition

Obtain medical data & Evaluate usability

Medical application

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

국제 연구협력



- UNIST XAI 연구센터 (한국, 울산/대전/서울)
- DARPA XAI 프로젝트 (미국, 워싱턴)
- 병원 네트워크 (중국, 상해/강소성/홍콩)
- Layer-Wise Relevance Propagation (독일, 베를린)
- Automatic Statistician (영국, 캠브리지)

미국 – DARPA XAI PI 미팅 참석(2019년 2월)

Gunning, David <david.gunning@darpa.mil> Wed, Oct 17, 2018 at 10:35 PM
To: Jaesik Choi <jaesik@unist.ac.kr>
Cc: "Carrera, Marisa (contr-i2o)" <marisa.carrera.ctr@darpa.mil>, "sw.lee" <sw.lee@korea.ac.kr>, "이경은 (전기전자컴퓨터공학부)" <socool@unist.ac.kr>

Jaesik,

Thank you very much for the opportunity. I was an honor to participate.

I've asked Marisa Carrera to add you to the list of people interested in XAI and to invite you to our February XAI Meeting at UC Berkeley.

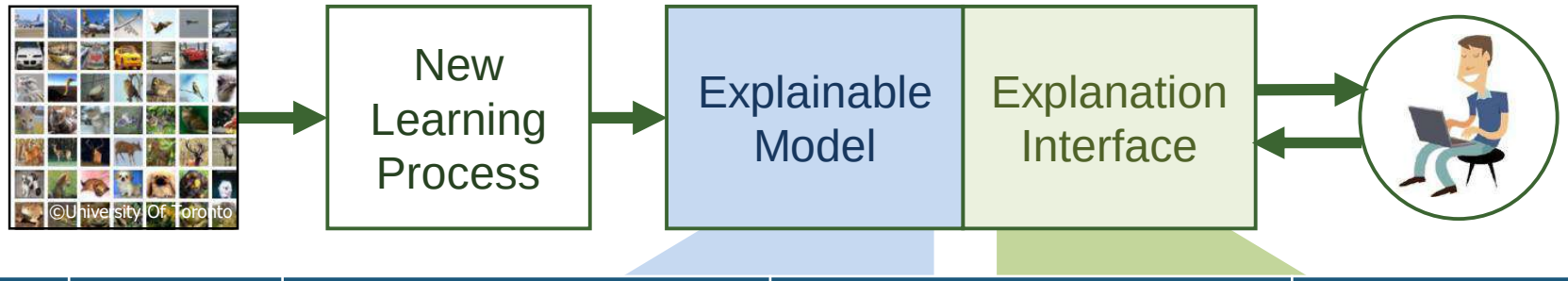
Dave

중국 – CDSS 이식/조인트벤처 추진(AITRICS)

상해/강소성 병원 네트워크
- CDSS 이식 및 조인트 벤처 설립

홍콩/광동성/북경 병원 네트워크
- CDSS 이식을 위한 협의 중

XAI Performers & Technical Approaches



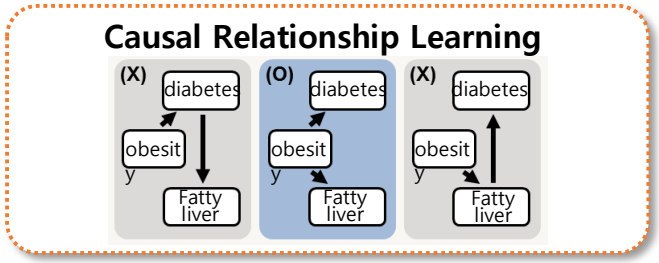
Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning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

* Korea: Korea University

R&D Objectives and co-PIs

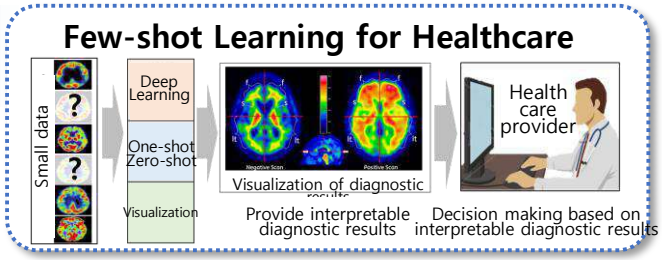
TA1 Learning Sparse Explainable Causal Features

KAIST - Sungju Hwang, Enho Yang and Jinwoo Shin



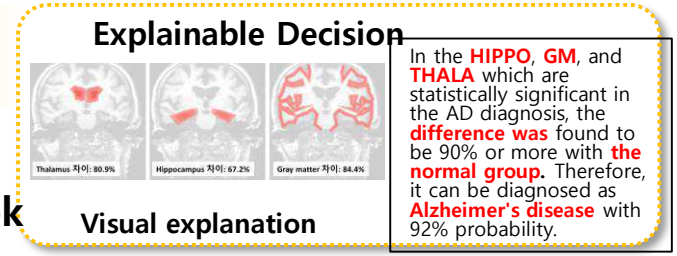
TA2 Few-shot Learning for Explainable Learners

KAIST - Hyunwook Park, Yongman Noh and Alice Oh
SNU - Bohyung Han
Yonsei - Seung-Won Hwang



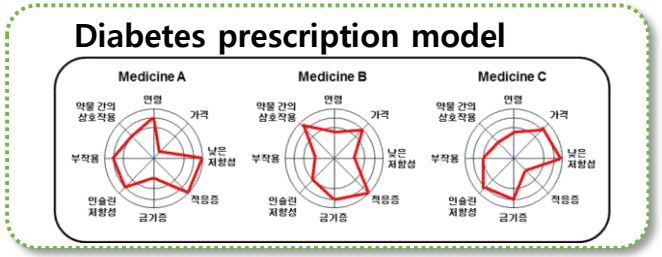
TA3 Develop Explainable Interface

UNIST - Jaesik Choi
Korea - Seong-Whan Lee, Klaus Muller and Heung-il Seok
KAIST - Kee-Eung Kim



TA4 Real-world application of Explainable AI

AITRICS – Jingyu Yu
Yeonsei (Hospital) – Kwangjun Kim
UNIST (Finance) – Byunggi Seo



Participating Researchers

Task 1
Explainable Learning for Causal Relations


Sungju Hwang


Enho Yang


Jinwoo Shin

Task 2
Few-Shots Learning for Explainable Learners


Hyunwook Park


Yongman Noh


Alice Oh


Bohyung Han


Seung-Won Hwang

Task 3
Explainable Interface


PI: Jaesik Choi


Seong-Whan Lee

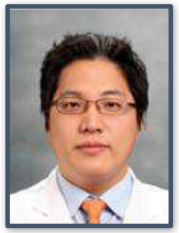

Klaus-Robert Muller


Heung-il Seok


Kee-Eung Kim

Task 4
Applications of Explainable Systems (Healthcare, Finance)


Jinkyu Yu


KwangJun Kim


Byunggi Seo

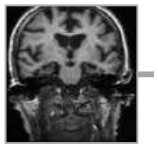
http://xai.unist.ac.kr/about/member/

XAI Framework

xaic
eXplainable Artificial Intelligence Center

Input

Healthcare



MR Scans
(900 patients)

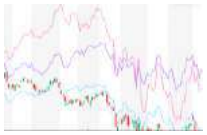


ICU hospital data
(58,000 patients)



National EHR data
(50.9M records)

Finance



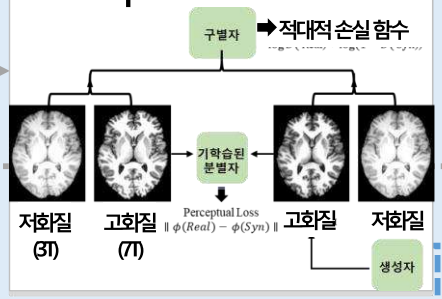
Reuter Stock



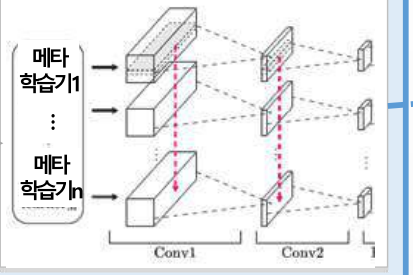
Koscom
SEC 10-K Reports
(about 60,000)

Task 2

Deep Generative Models

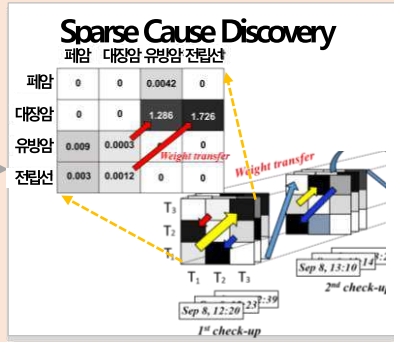


Few Shot Learning

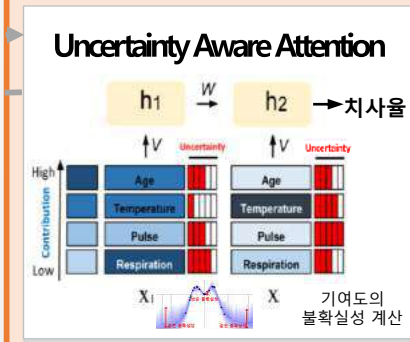


Task 1

Sparse Cause Discovery

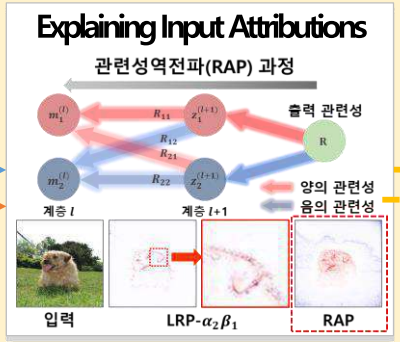


Uncertainty Aware Attention

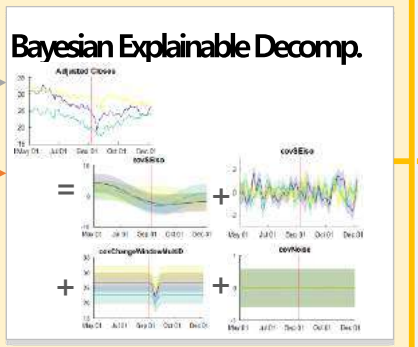


Task 3

Explaining Input Attributions



Bayesian Explainable Decomp.



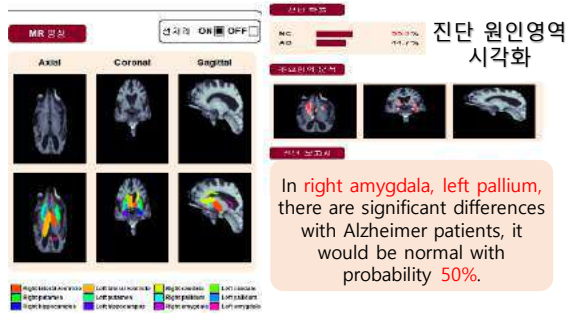
Output

Task 4

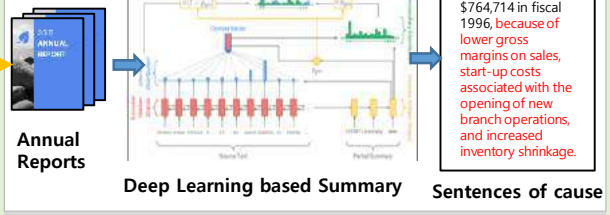
Explainable Medical Diagnosis



Explainable Interface for Alzheimer



Explaining Financial Data



Technical Approaches

Explaining Input Attributions of Deep Neural Networks

Explaining Deep Neural Networks w/ Sparse Features

Few Shots Learning for Visual Question Answering

Explainable Multivariate Nonparametric Bayesian Models

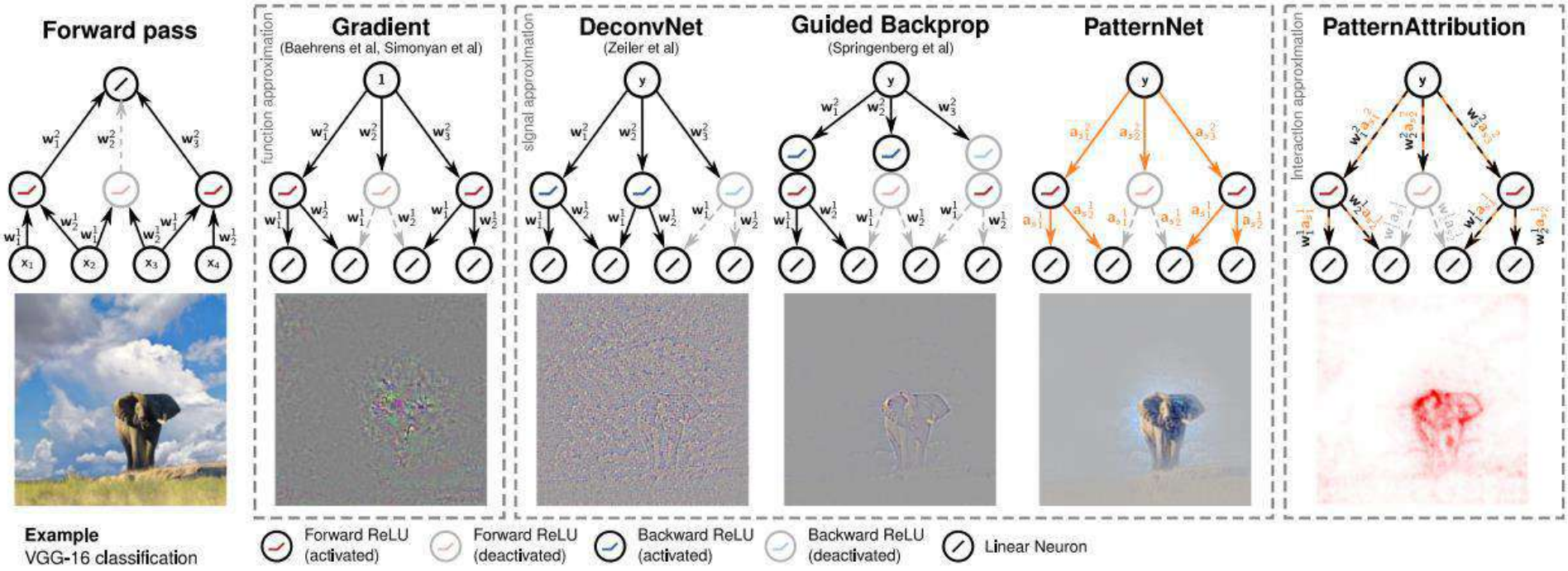
Explainable Interface of Deep Reinforcement Learning

Explaining Medical Diagnosis with Adversarial Training

PatternNet – Learning to Explain Neural Net – 1/2

Klaus-Robert Müller (TU Berlin/Korea University)

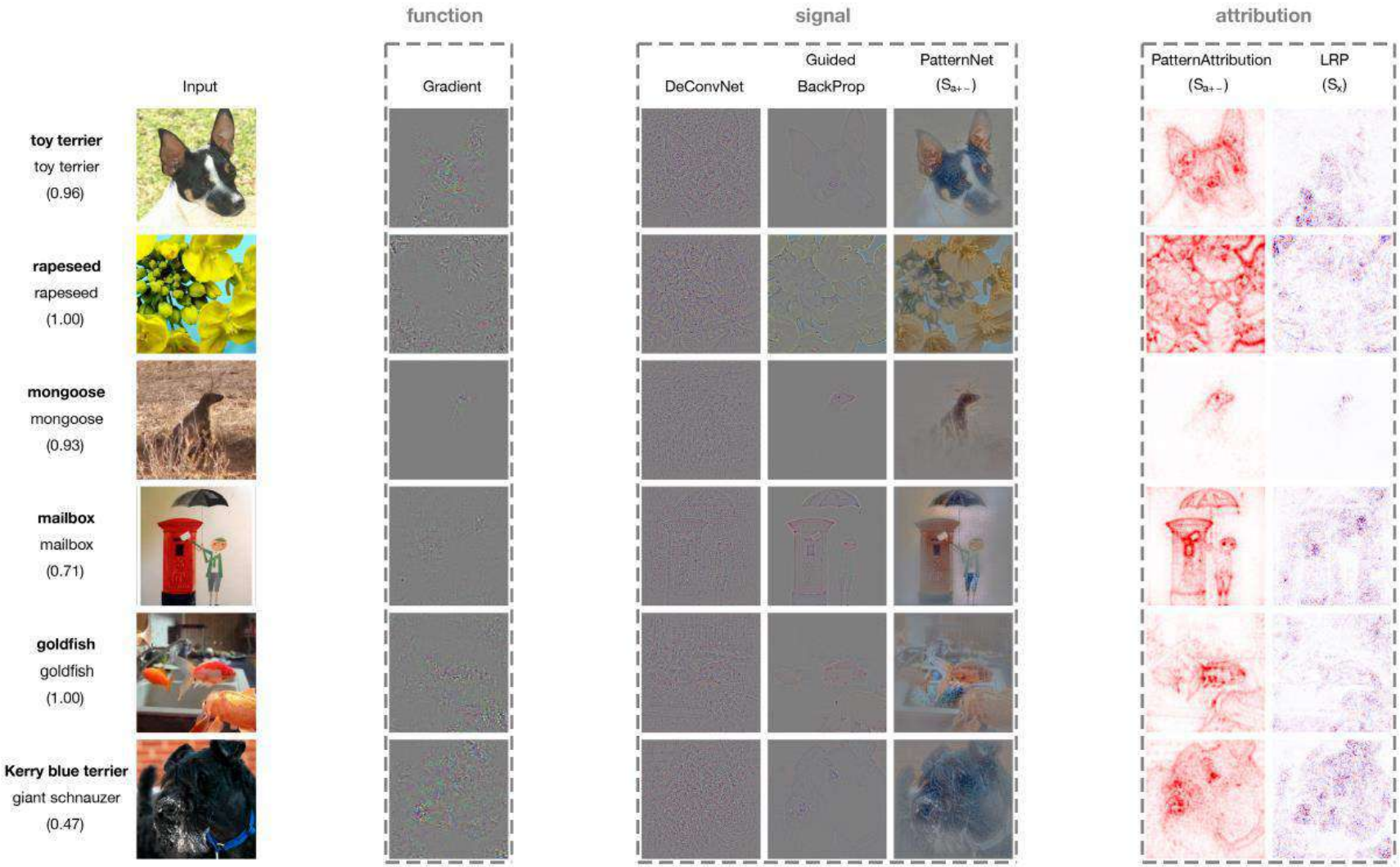
Learning (Optimizing) Attributions to Reduce Signal Noise



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

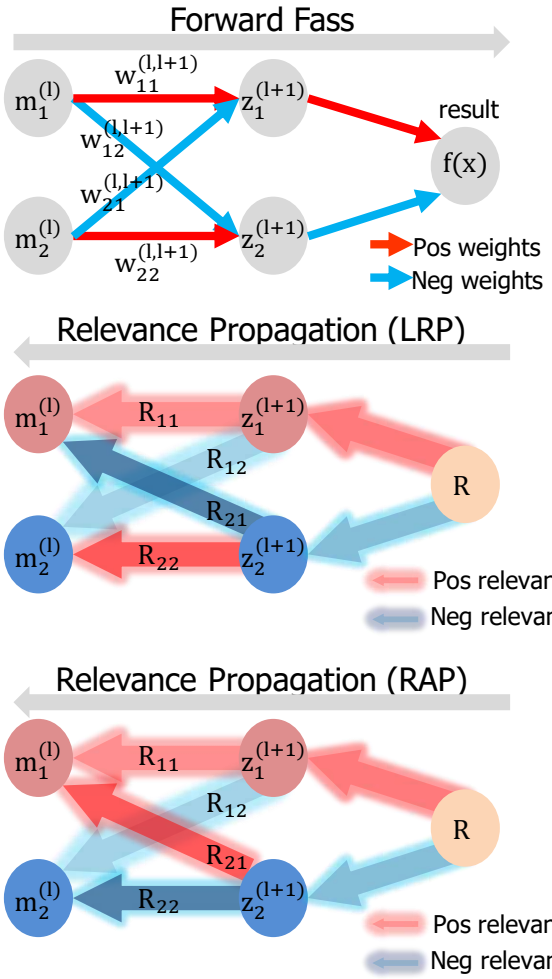
Klaus-Robert Müller (TU Berlin/Korea University)



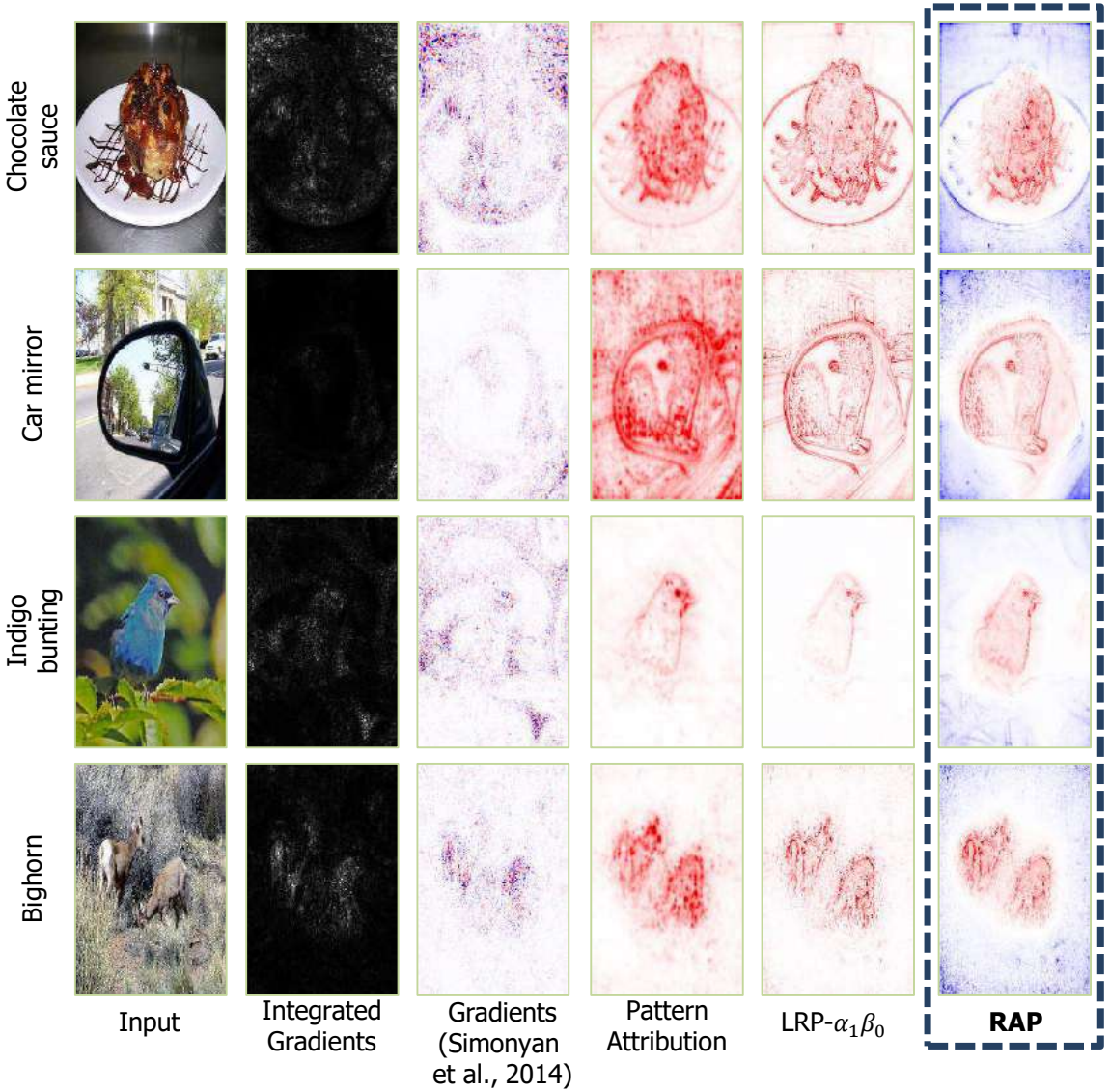
Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018

RAP - Relative Attribution Propagation

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



Propagating flipped attributions of negative relevance



Woo-Jeong Nam, Seong-Whan Lee and Jaesik Choi, "Relative Attribution Propagation", soon will be available at arxiv

Explaining Medical Diagnosis of Brain MR Images

Seong-Whan Lee (Korea University)



Technical Approaches

Explaining Input Attributions of Deep Neural Networks

Explaining Deep Neural Networks w/ Sparse Features

Few Shots Learning for Visual Question Answering

Explainable Multivariate Nonparametric Bayesian Models

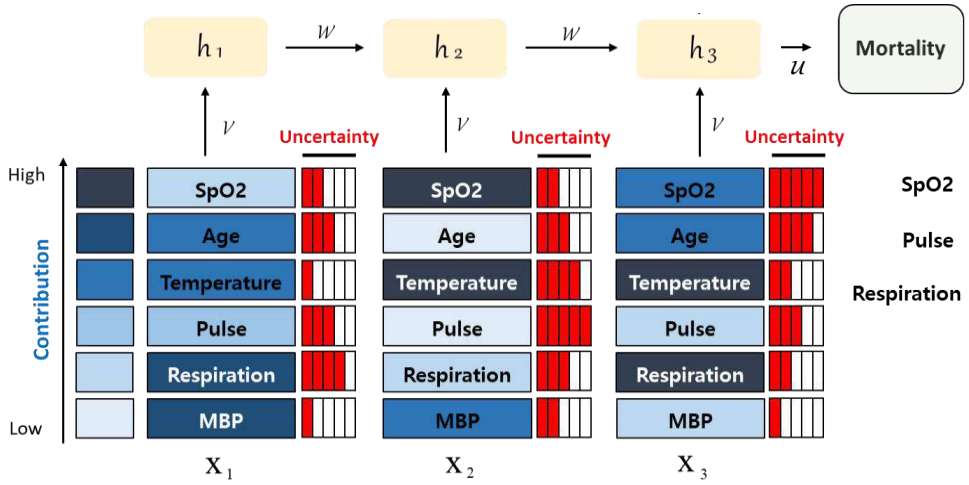
Explainable Interface of Deep Reinforcement Learning

Explaining Medical Diagnosis with Adversarial Training

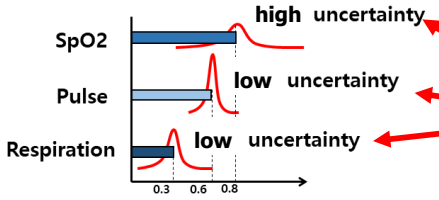
Uncertainty Aware Attention – 1/2

Sung Ju Hwang & Eunho Yang (KAIST, AItrics)
Kwangjun Kim (Yeonsei Hospital)

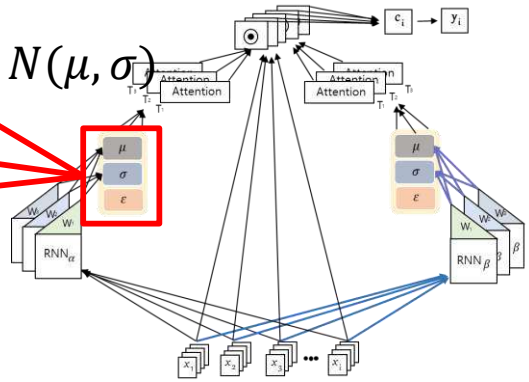
Explaining the uncertainty of features selected by attention models



Uncertainty Aware Attention (UA) Model



Variances of attentions on features



RETAIN + Uncertainty for Attention

Solutions: Dropout + Bayesian Inference + Computing confidence on individual data

- UA models outperforms existing uncertainty aware RETAIN models on PhysioNet, Pancreatic Cancer and MIMIC-Sepsis

	PhysioNet				Pancreatic Cancer	MIMIC Sepsis
	Mortality	Stay < 3	Cardiac	Recovery		
RETAIN-DA [3]	0.7652± 0.02	0.8515± 0.02	0.9485± 0.01	0.8830± 0.01	0.8528± 0.01	0.7965± 0.01
RETAIN-SA [28]	0.7635± 0.02	0.8412± 0.02	0.9360± 0.01	0.8582± 0.02	0.8444± 0.01	0.7695± 0.02
UA-Independent	0.7764± 0.01	0.8572± 0.02	0.9516± 0.01	0.8895± 0.01	0.8533± 0.03	0.8019± 0.01
UA	**0.7827± 0.02**	**0.8628± 0.02**	0.9563± 0.01	0.9049± 0.01	0.8604± 0.01	0.8017± 0.01
UA+	0.7770± 0.02	0.8577± 0.01	**0.9612± 0.01**	**0.9074± 0.01**	**0.8638±0.02**	**0.8114± 0.01**

Jay Heo, Haebeom Lee, Saehoon Kim, Juho Lee, Kwangjun Kim, Eunho Yang, and Sung Ju Hwang, "Uncertainty-Aware Attention for Reliable Interpretation and Prediction", NeurIPS 2018

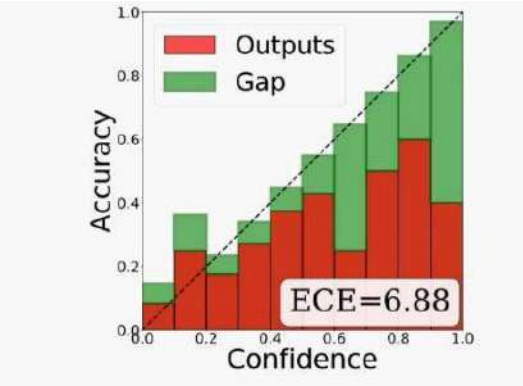
49

Uncertainty Aware Attention – 2/2

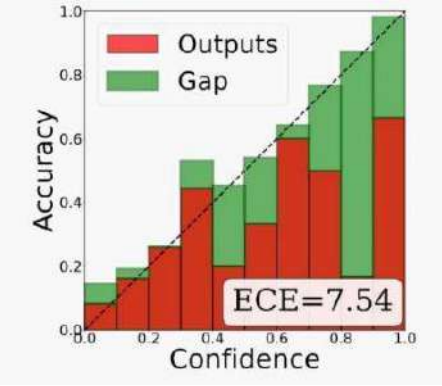
Sung Ju Hwang & Eunho Yang (KAIST, AItrics)
Kwangjun Kim (Yeonsei Hospital)

Explaining the uncertainty of features selected by attention models

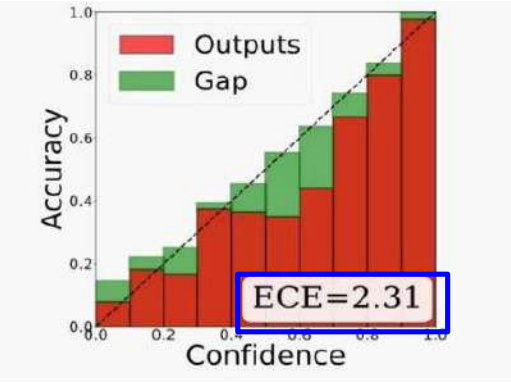
In terms of Expected Calibration Error (ECE), UA provides improved confidence/accuracy match on explaining uncertainty of features compared to RETAIN-DA (Deterministic Attention) and RETAIN-SA (Stochastic Attention).



RETAIN-DA (Choi et al., 2016)



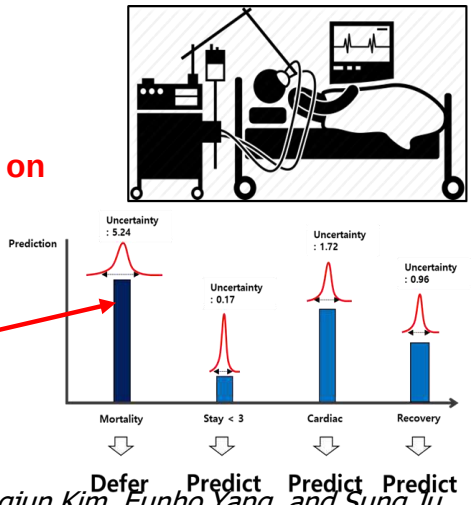
RETAIN-SA (Xu et al., 2015)

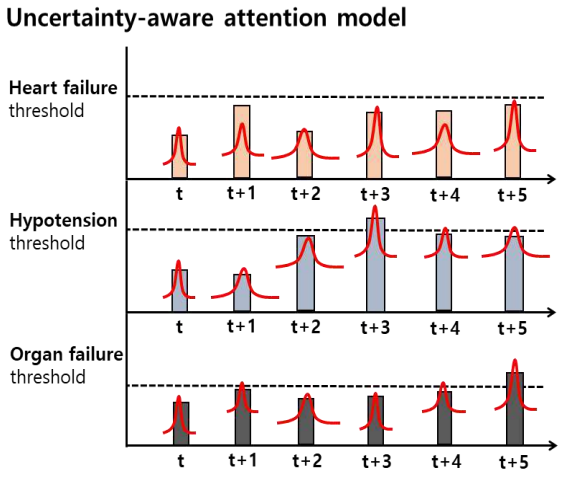


UA

Computing the prediction uncertainty based on feature uncertainty (of attention)

For outputs with high variance, It says 'I Don't Know'



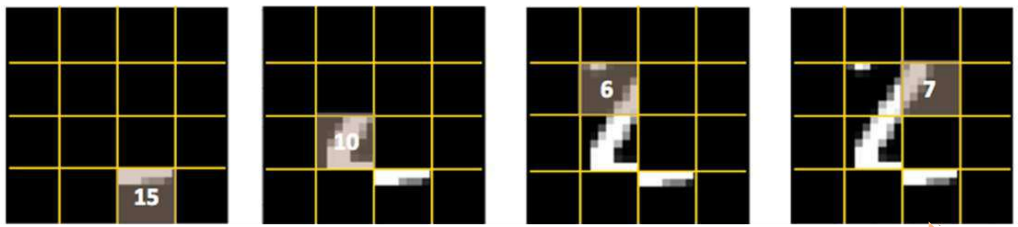


Jay Heo, Haebeom Lee, Saehoon Kim, Juho Lee, Kwangjun Kim, Eunho Yang, and Sung Ju Hwang, "Uncertainty-Aware Attention for Reliable Interpretation and Prediction", NeurIPS 2018

Reinforcement Learning for Sequential Features – 1/2

Eunho Yang & Sung Ju Hwang (KAIST, AItrics)

Explaining an inexpensive and accurate order of feature acquisition

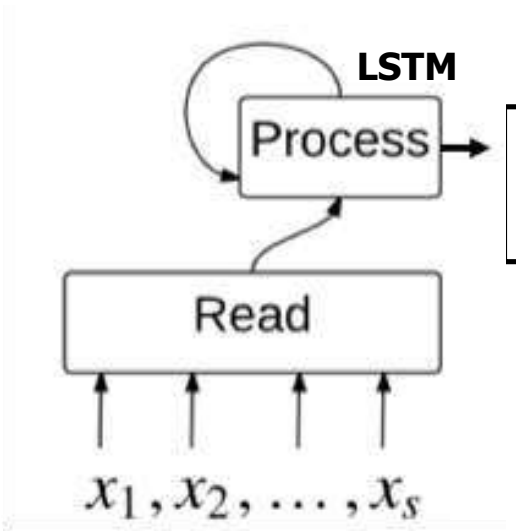




Result Is 2!

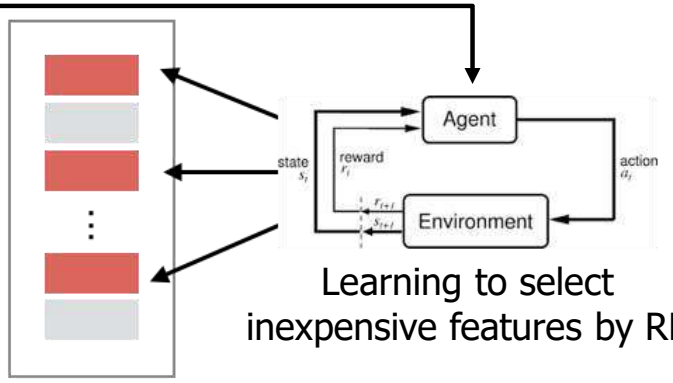
An order of feature acquisition by RL

Solutions: Joint learning of Set Encoding + Reinforcement Learning (Q Learning)

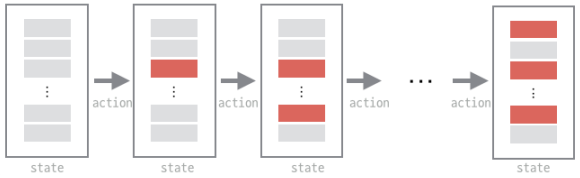


an encoding of a subset (x1, …, xs)

Set Encoding (Vinyals et al., 2016)



Learning to select inexpensive features by RL



Hajin Shim, Sung Ju Hwang and Eunho Yang, "Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding", NeurIPS 2018

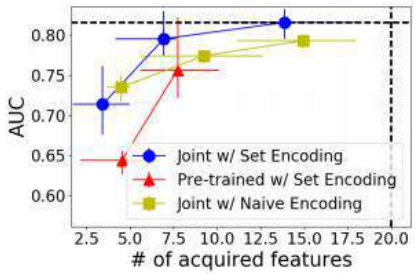
Reinforcement Learning for Sequential Features – 2/2

Eunho Yang & Sung Ju Hwang (KAIST, AItrics)

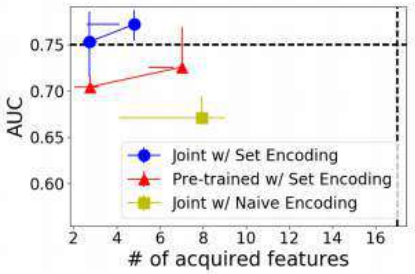
Explaining an inexpensive and accurate order of feature acquisition

In terms of AUC and acquisition cost, 'Joint w/ Set Encoding' provides competitive results by significant reduction of cost.

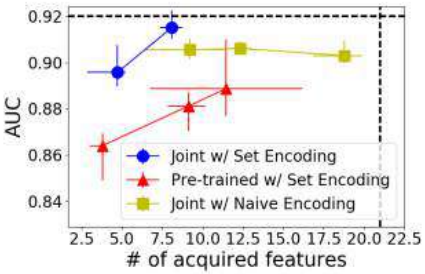
Physionet 2012 – ICU (Intensive Care Unit) dataset



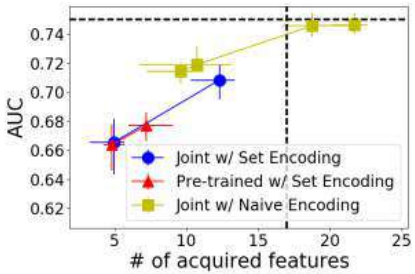
(a) Mortality



(b) Length of stay < 3



(c) Cardiac condition



(d) Recovery from surgery

Medical checkup

AUC

model / task	Baselines	Joint w/ Naive Encoding		Pre-trained w/ Set Encoding		Joint w/ Set Encoding	
	MLP	uniform cost	real cost	uniform cost	real cost	uniform cost	real cost
Fatty Liver	0.812 35 (577.7)	0.775 6.81 (167.0)	0.785 8.7 (20.1)	0.774 6.2 (127.9)	0.779 4.8 (9.2)	0.795 10.3 (149.8)	0.792 7.5 (13.8)
IMT	0.755 22 (1,629.8)	0.742 5.12 (109.9)	0.704 14.22 (11.3)	0.721 3.94 (191.5)	0.760 23.71 (83.8)	0.756 5.57(309.8)	0.750 4.79 (10.0)

intima-media thickness

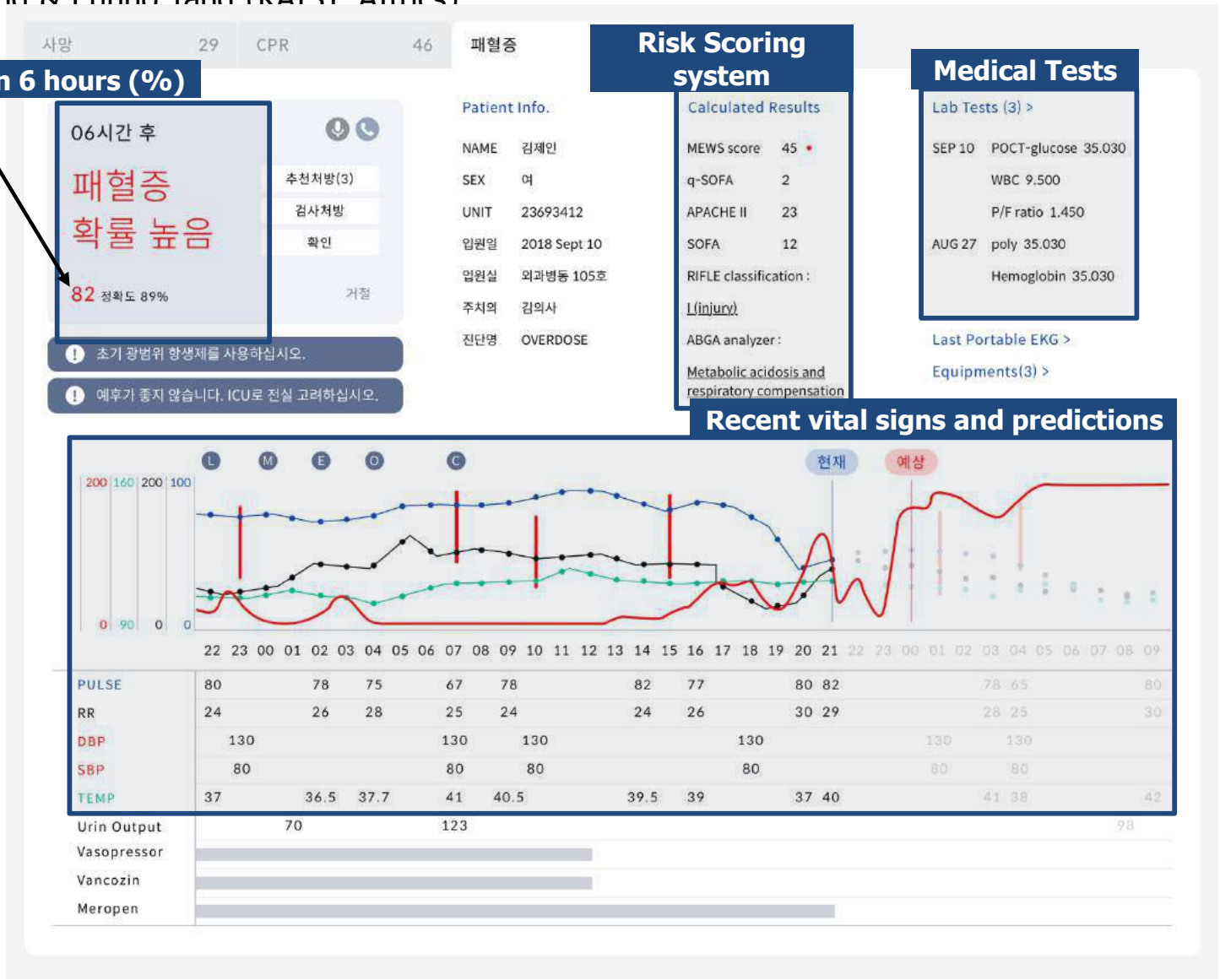
# of features

cost in USD

Hajin Shim, Sung Ju Hwang and Eunho Yang, "Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding", NeurIPS 2018

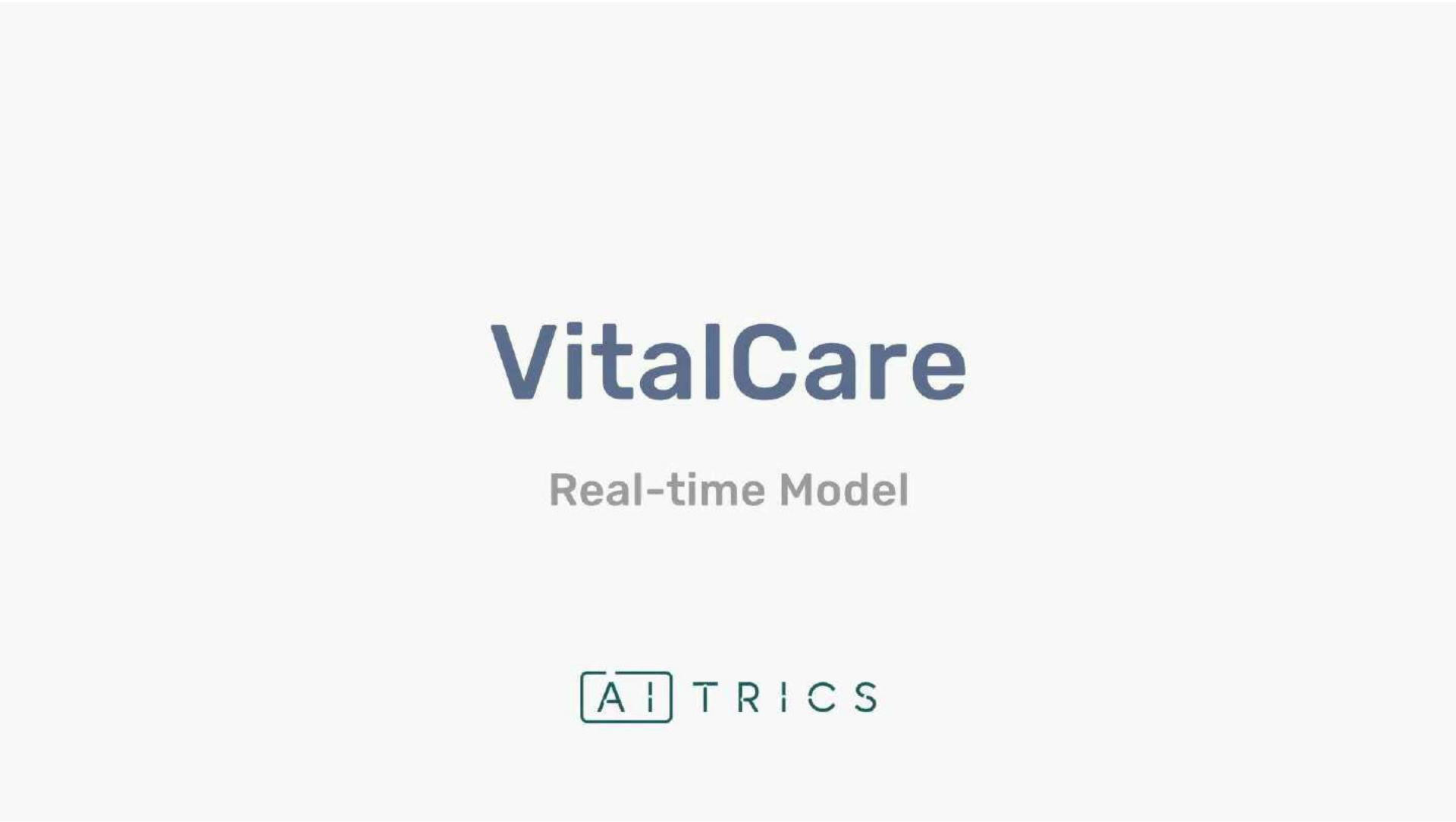
VitalCare: Explaining Risks of Patients in ICU

Jinkyu Yu (AItrics), Kwangjun Kim (Yeonsei Hospital)
Sung Ju Hwang & Eunho Yang (KAIST, AItrics)



VitalCare: Explaining Risks of Patients in ICU

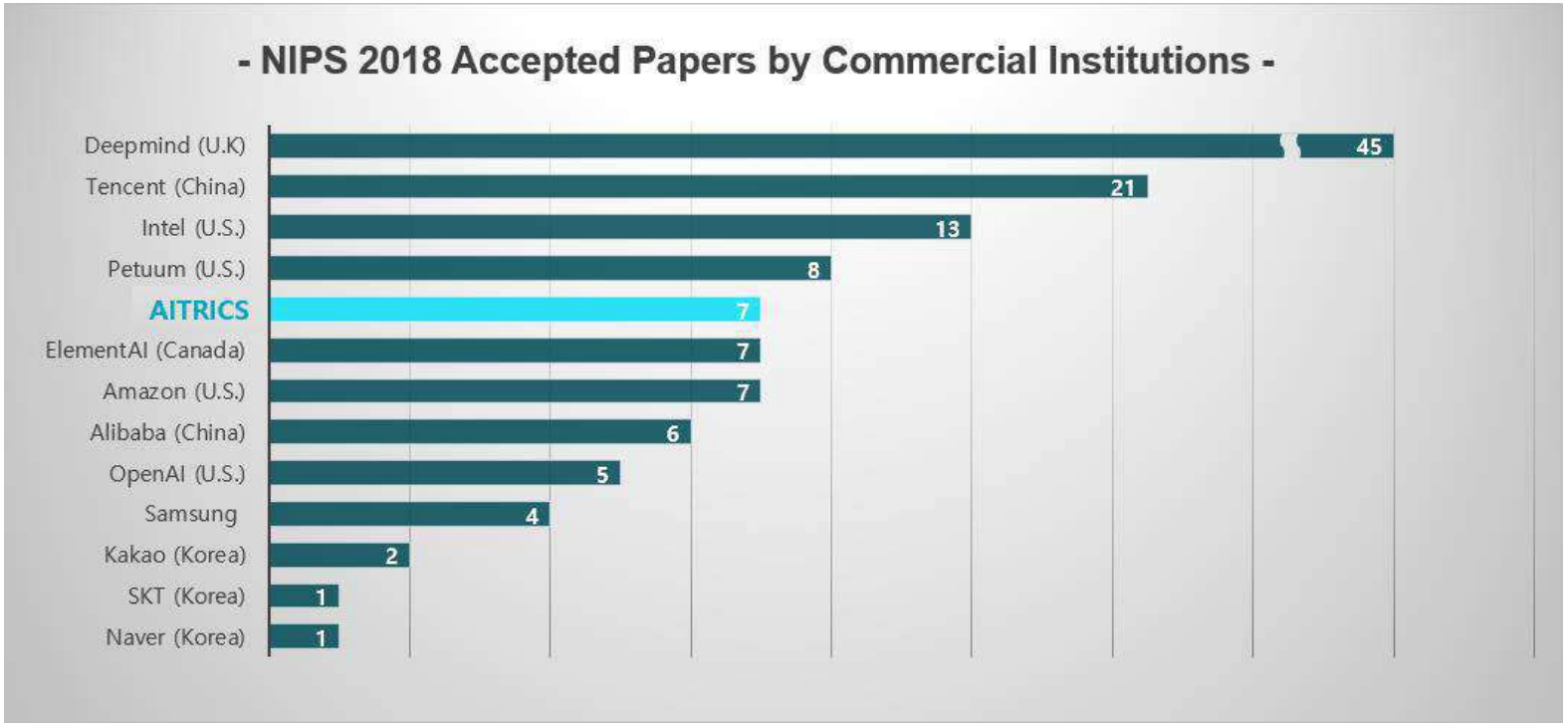
Jinkyu Yu (AItrics), Kwangjun Kim (Yeonsei Hospital)
Sung Ju Hwang & Eunho Yang (KAIST, AItrics)



RiskCare: Explaining Risks in Medical Checkups

Jinkyu Yu (AItrics), Kwangjun Kim (Yeonsei Hospital)
Sung Ju Hwang & Eunho Yang (KAIST, AItrics)

AITRICS in NIPS 2018







Technical Approaches

Explaining Input Attributions of Deep Neural Networks

Explaining Deep Neural Networks w/ Sparse Features

Few Shots Learning for Visual Question Answering

Explainable Multivariate Nonparametric Bayesian Models

Explainable Interface of Deep Reinforcement Learning

Explaining Medical Diagnosis with Adversarial Training