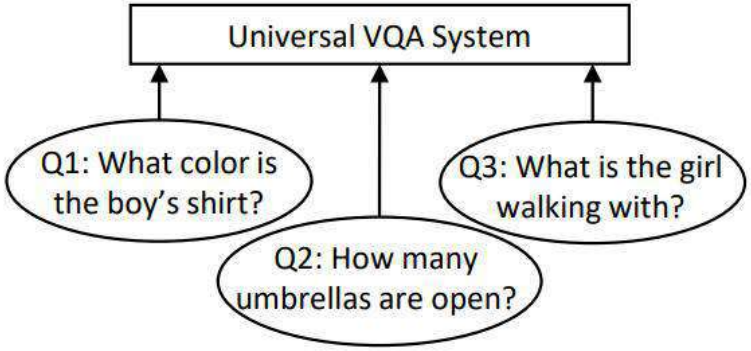
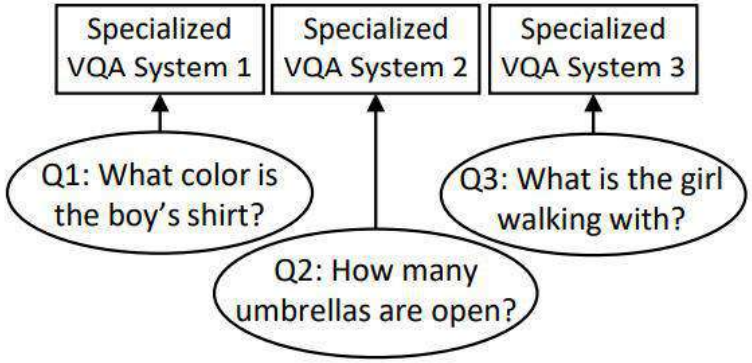
Using General Knowledge for Visual Q&A – 1/2

Bohyung Han (SNU)
Jinwoon Shin (KAIST)

Utilizing General Knowledge (over all training data) to train specialized leaners

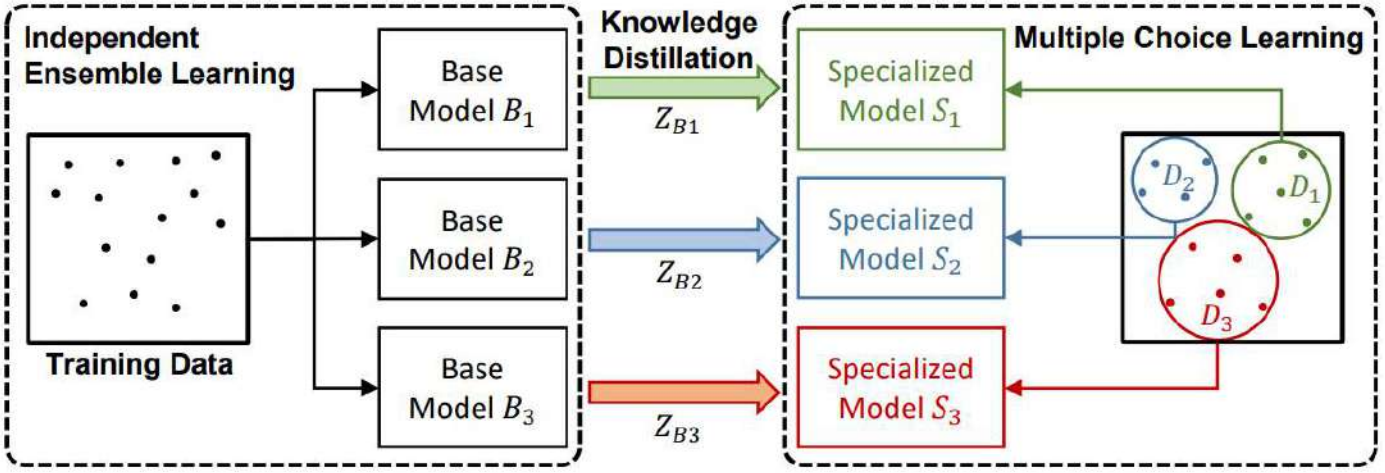


A Conventional Ensemble



Multiple Choice Learning (MCL) (Guzman-Rivera et al., 2012)

Problem of data deficiency of MCL is solved by Knowledge Distillation from General Base Learners



Jonghwan Mun, Kimin Lee, Jinwoo Shin, Bohyung Han, "Learning to Specialize with Knowledge Distillation for Visual Question Answering", NeurIPS 2018

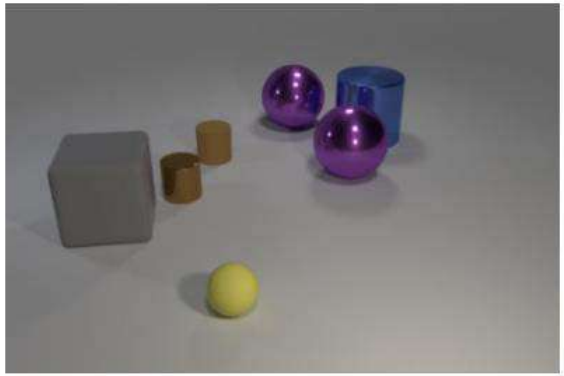
Using General Knowledge for Visual Q&A – 2/2

Bohyung Han (SNU)
Jinwoon Shin (KAIST)

Utilizing General Knowledge (over all training data) to train specialized leaners

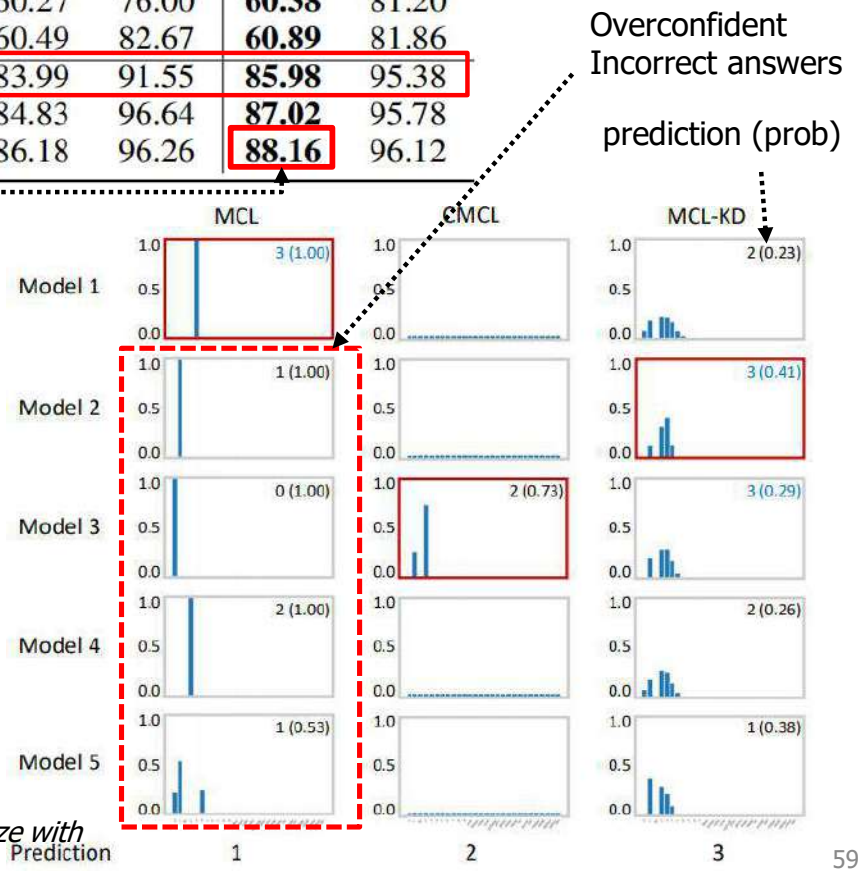
CLEVER dataset

Answering Network	k	Single Top1	IE Top1	IE Oracle	MCL Top1	MCL Oracle	CMCL Top1	CMCL Oracle	MCL-KD Top1	MCL-KD Oracle
MLP	1	58.40	60.10	80.73	41.31	98.92	59.12	63.76	**60.22**	80.75
	2				48.94	97.57	60.27	76.00	**60.38**	81.20
	3				58.63	95.67	60.49	82.67	**60.89**	81.86
SAN	1	82.30	85.23	94.93	42.19	98.67	83.99	91.55	**85.98**	95.38
	2				58.39	98.62	84.83	96.64	**87.02**	95.78
	3				83.73	98.62	86.18	96.26	**88.16**	96.12



Q: what number of objects are either big objects that are behind the big gray block or tiny brown rubber balls?

Ground Truth Answer: 3



Jonghwan Mun, Kimin Lee, Jinwoo Shin, Bohyung Han, "Learning to Specialize with Knowledge Distillation for Visual Question Answering", NeurIPS 2018

Machine-Translated Knowledge Transfer for Causal Reasoning

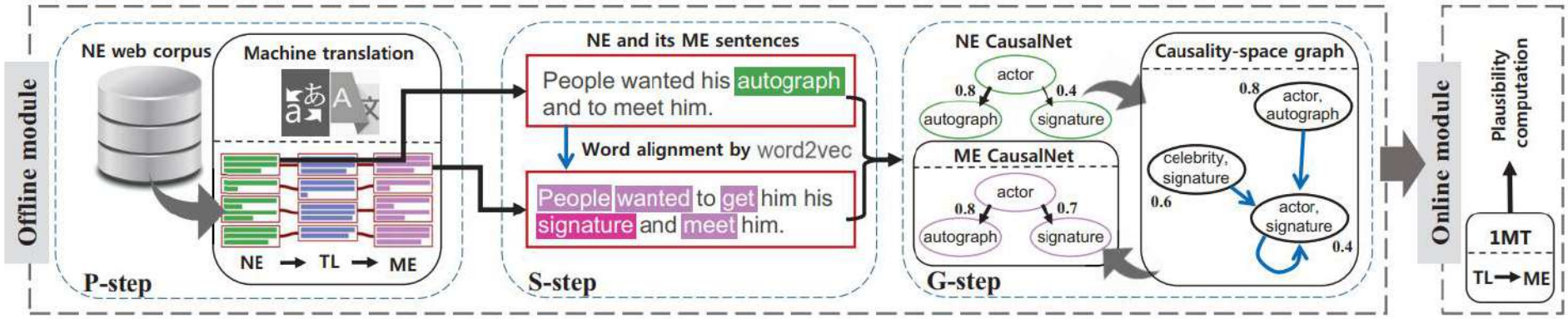
Seung-won Hwang (Yonsei)

Utilizing multi-lingual data to improve Causal Reasoning

Choice of Plausible Alternatives (COPA) evaluation (Roemmele et al., 2011)

Example 1
Premise: The girl met her favorite actor. What happened as an effect?
Alternative 1: She went to see his new film.
Alternative 2: She asked him for his autograph

Problem of data deficiency in Causal Reasoning is solved by Machine Translation
(Knowledge Distillation from Machine Translation Engine)



Pseudo corpus generation **Selective word alignment** **Graph transformer**

Method	Korean	Chinese	French
TransCP	.568 (80.9%)	.572 (81.5%)	.560 (79.8%)
TransCN	.574 (81.8%)	.580 (82.6%)	.568 (80.9%)
TransQA	.642 (91.5%)	.634 (90.3%)	.604 (86.0%)
TransQA+	.546 (77.9%)	.570 (81.2%)	.580 (82.6%)
PSG	**.690** (98.3%)	**.662** (94.3%)	**.662** (94.3%)

Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang, Machine-Translated Knowledge Transfer for Commonsense Causal Reasoning, AAAI 2018

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

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

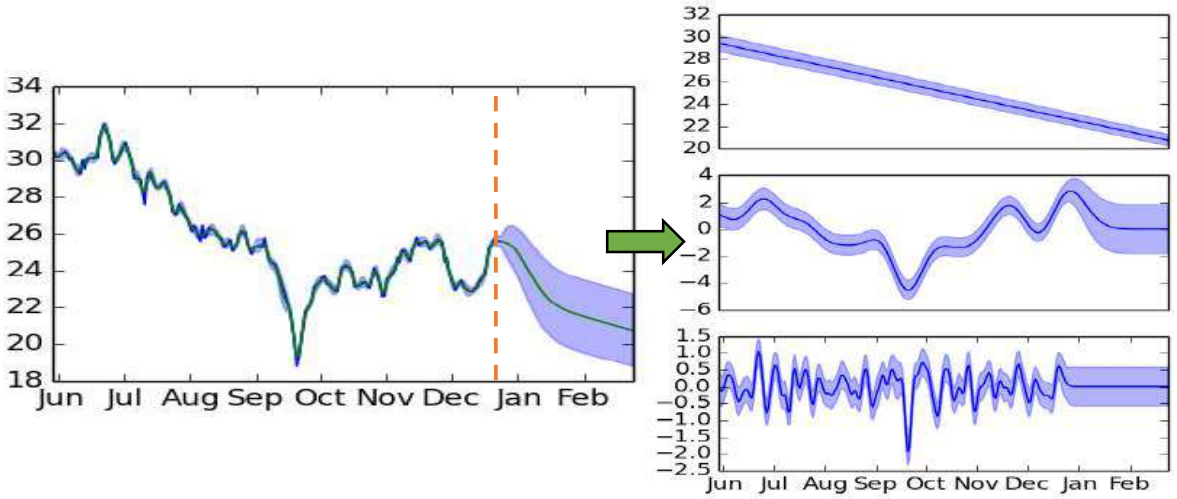
Explaining Medical Diagnosis with Adversarial Training

Relational Automatic Statistician – 1/3

Jaesik Choi (UNIST)

Explaining Multivariate Time Series in Financial Prediction

The Automatic Statistician System (Lloyd et al, 2014)

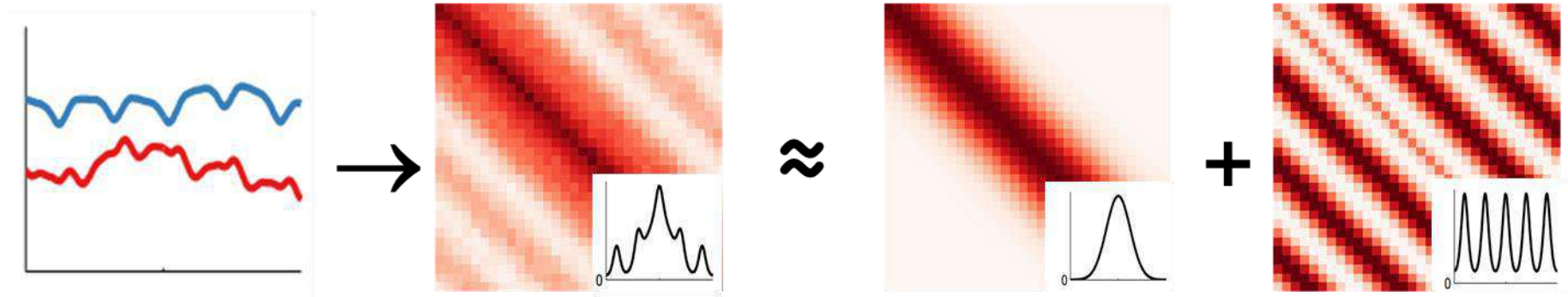


Linear function
decrease x/week

Smooth function
Length scale: y weeks

Rapidly varying
smooth function
Length scale: z hours

Covariance Decomposition: Learn Explainable Models from Data

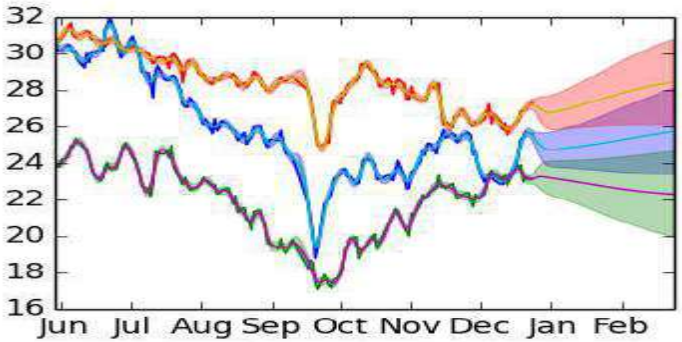


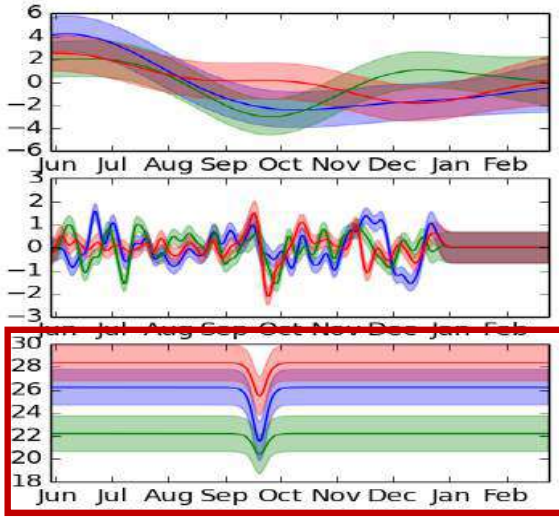
Relational Automatic Statistician – 2/3

Jaesik Choi (UNIST)

Explaining Multivariate Time Series in Financial Prediction

Relational Automatic Statistician System (Hwang et al, 2016)



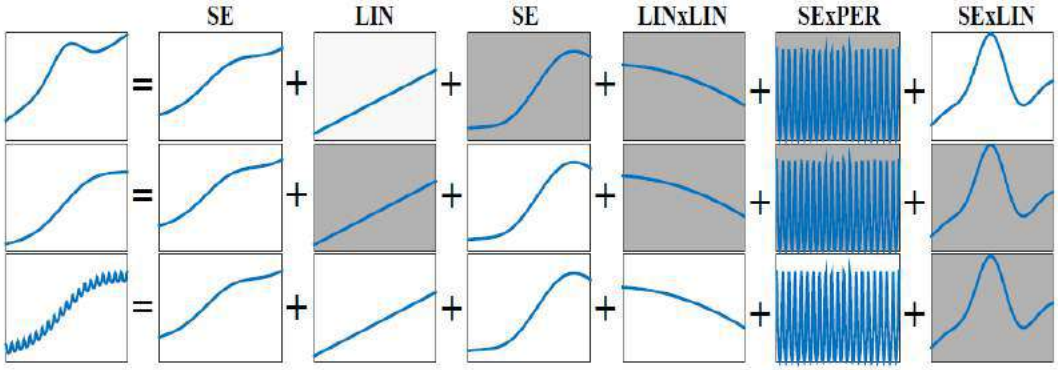


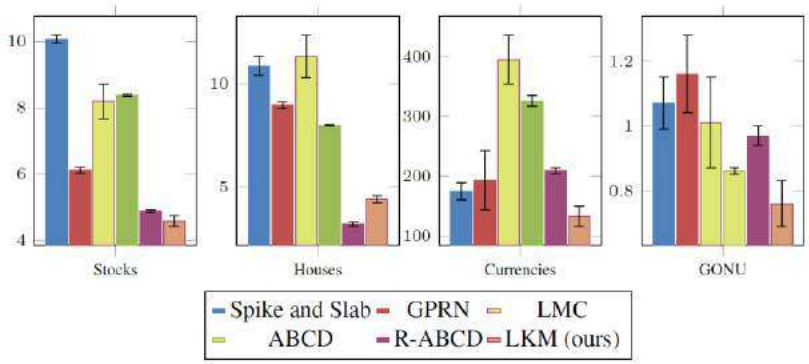
Smooth function
Length scale: y weeks

Rapidly varying
smooth function
Length scale: z hours

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

Latent Kernel Model





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", IJCAI 2018 workshop

Relational Automatic Statistician – 3/3

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:

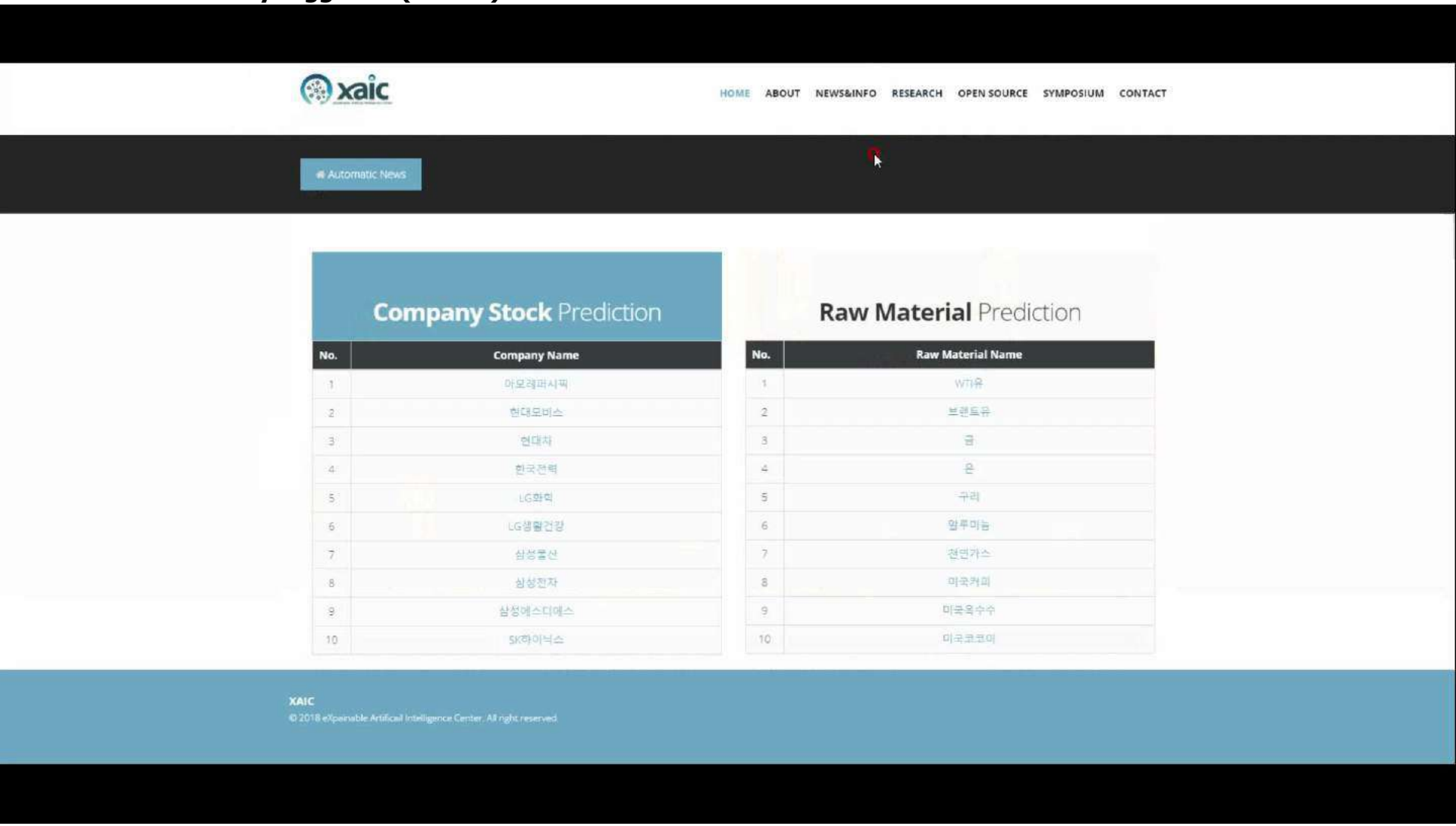
This component is a linear function.

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", IJCAI 2018 workshop

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", IJCAI 2018 workshop

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

Explainable Deep Reinforcement Learning – 1/2

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

Explainable Deep Reinforcement Learning for Continuous Action Space

Curling strategy requires precise action selection on a large action space

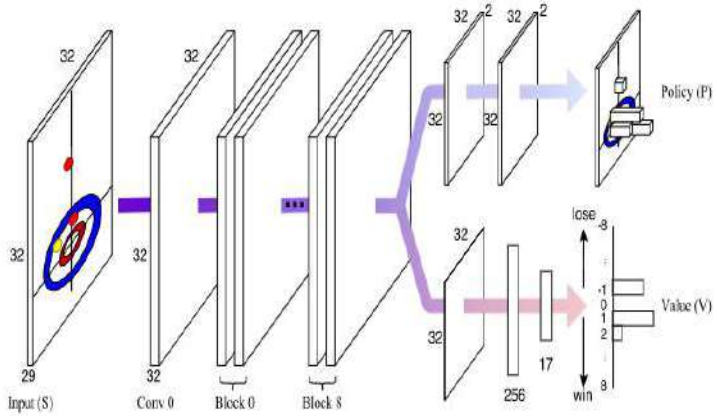


Kyowoon Lee, Sol-A Kim, Jaesik Choi, and Seong-Whan Lee, "Deep Reinforcement Learning in Continuous Action Spaces: a Case Study in the Game of Simulated Curling", ICML 2018

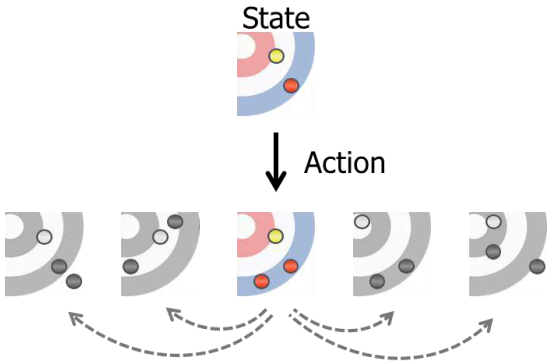
Explainable Deep Reinforcement Learning – 2/2

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

Explainable Deep Reinforcement Learning for Continuous Action Space

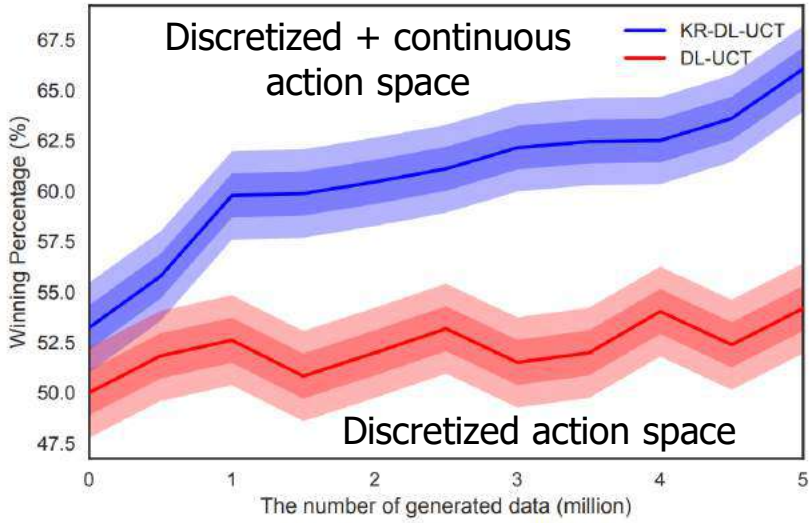


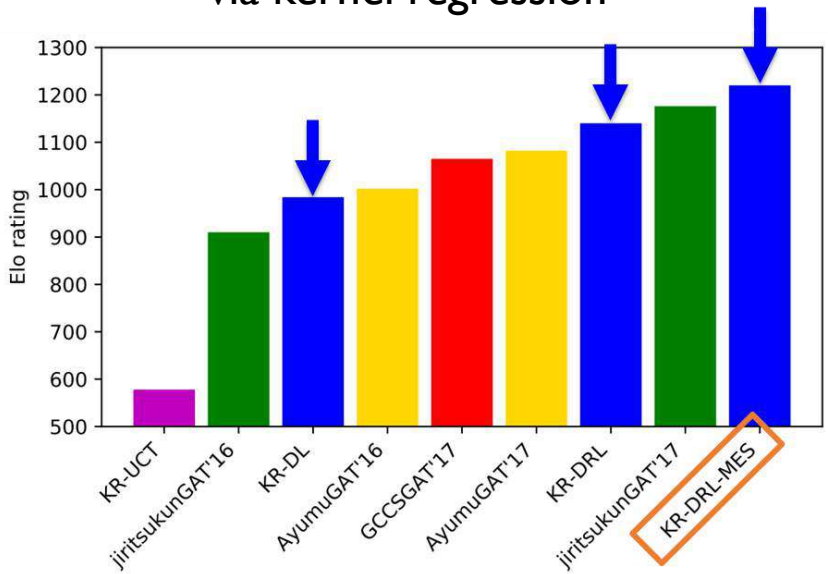
+



Shares information among similar actions via kernel regression

Deep Reinforcement Learning for Curling





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

Won in Game AI Tournament (GAT)-2018

Explaining Deep Reinforcement Learning – Demo

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



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

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

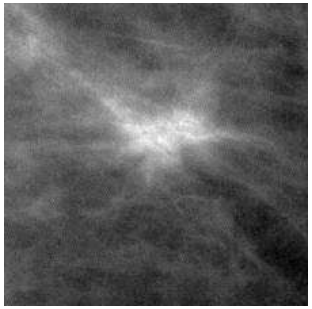
Explainable Interface of Deep Reinforcement Learning

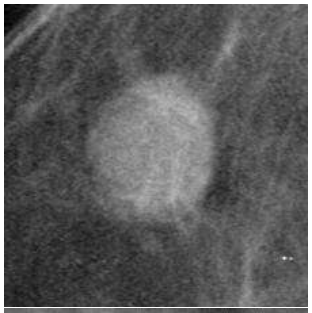
Explaining Medical Diagnosis with Adversarial Training

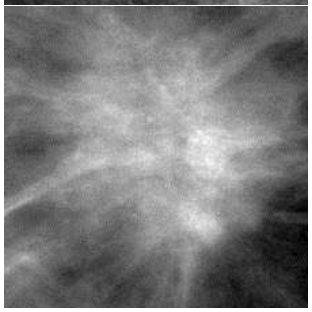
XCAD: Explainable Computer Aided Diagnosis – 1/2

Yong Man Ro (KAIST)

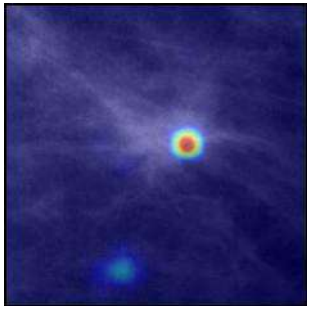
Test image

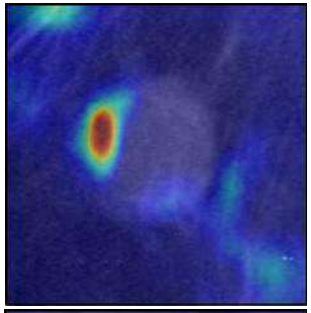


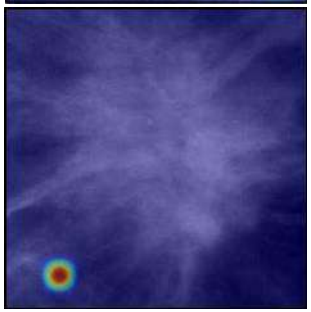




grad-CAM



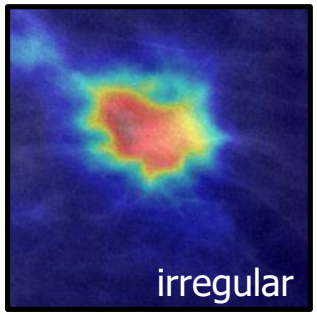


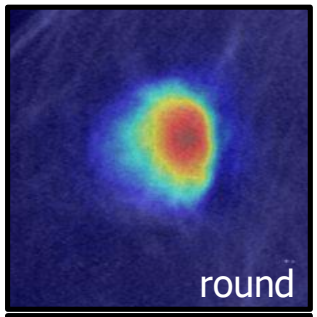


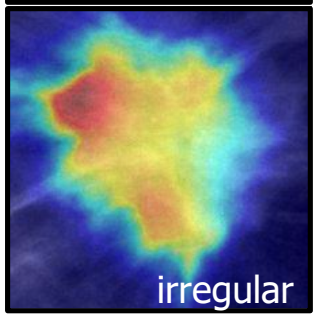
Difficult for medical doctors

Explanations of XCAD guided by BI-RADS*

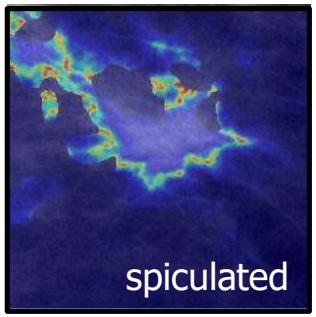
Shape

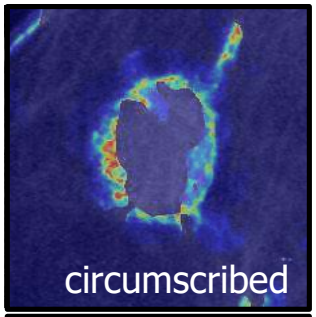


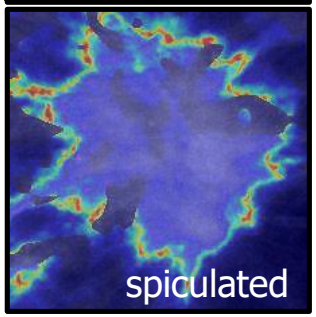




Margin







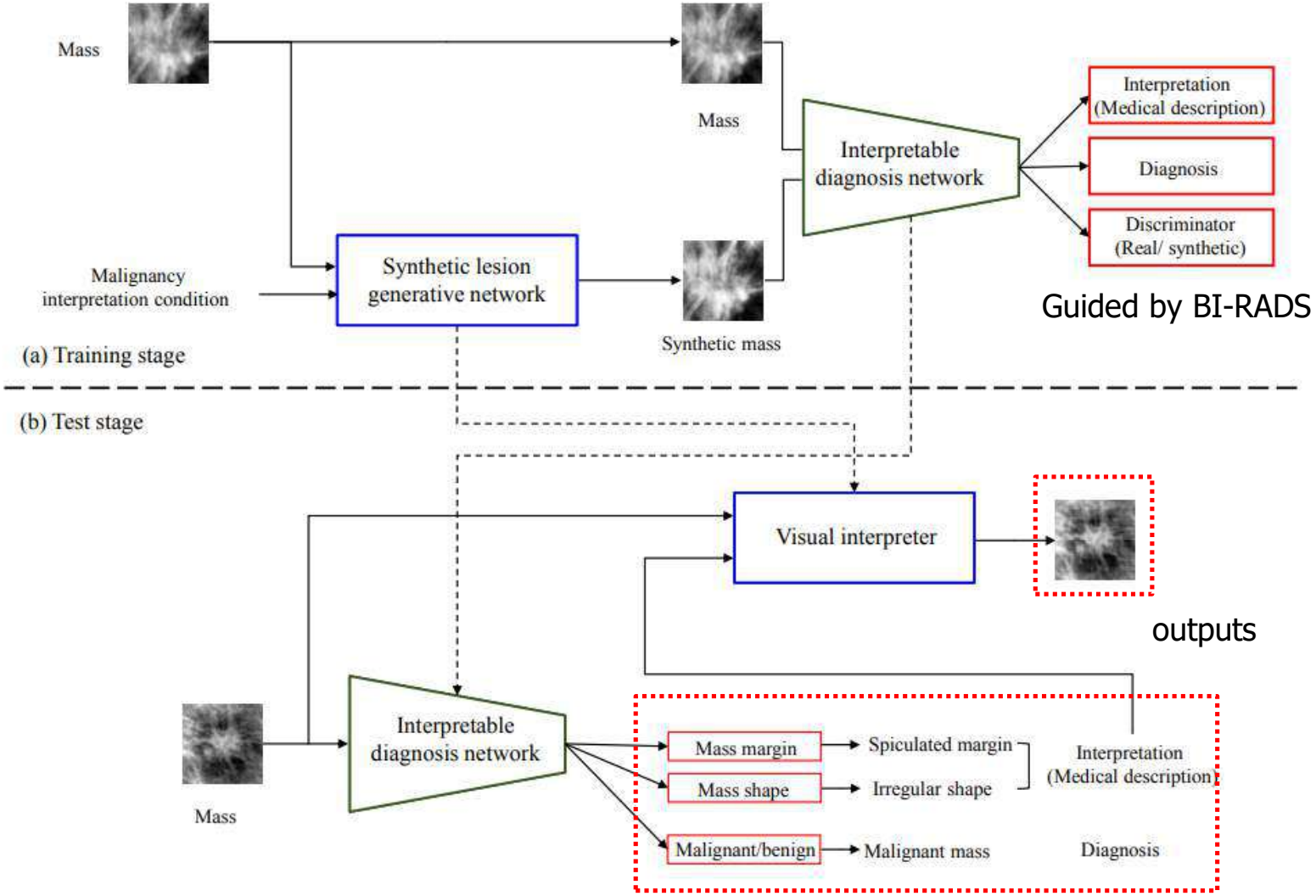
Explanations that are familiar to doctors

**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)

목차

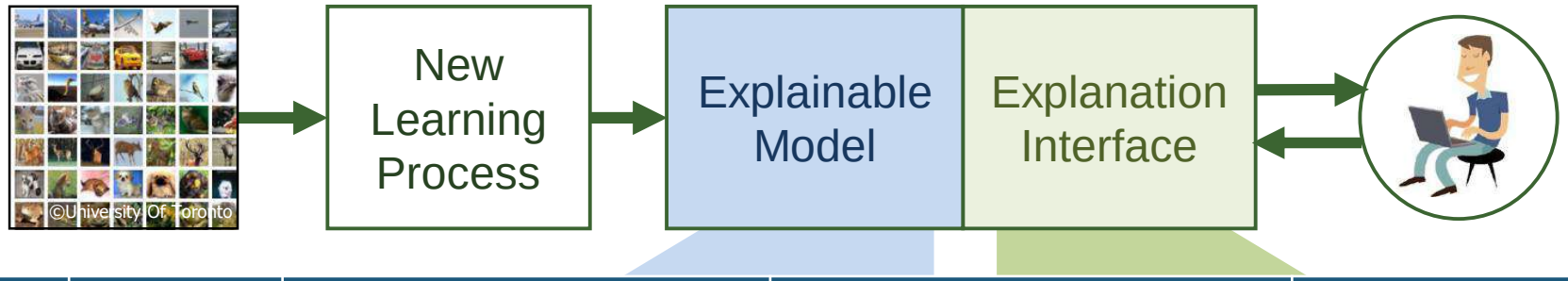
소개

미국의 DARPA의 XAI 프로젝트

과학기술정보통신부 설명가능인공지능 연구센터

향후 계획

Future Plans



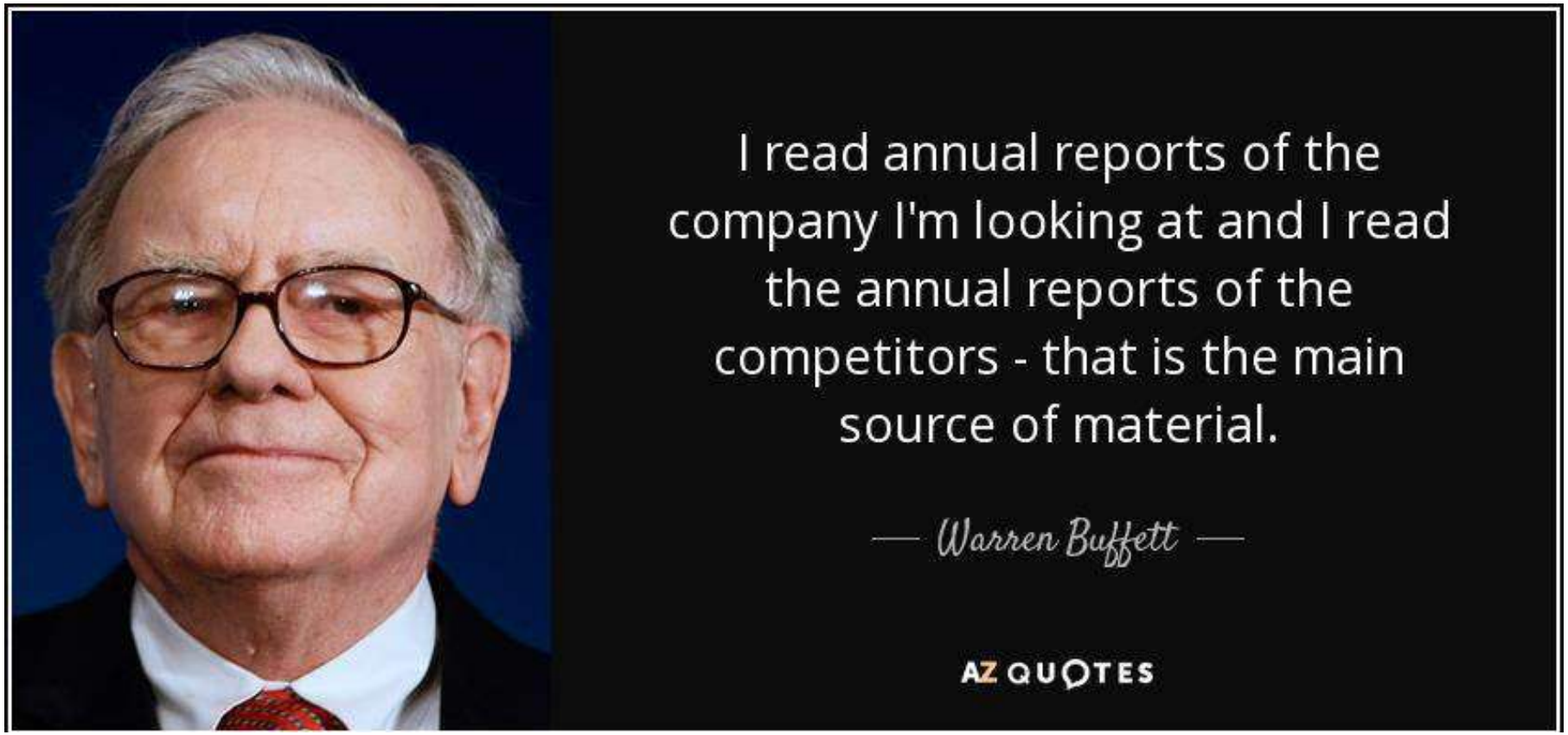
Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning	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

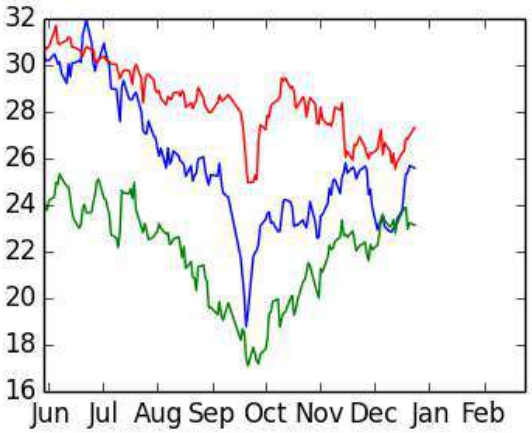
Future: Toward Reading/Explaining Reports

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.

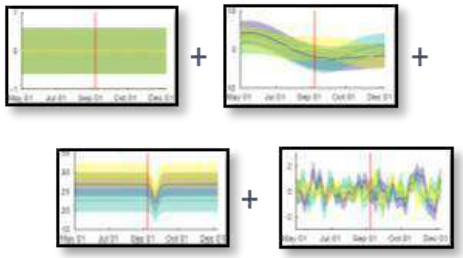
Future: Toward Reading/Explaining Reports



Future: Toward Reading/Explaining Reports



Multiple Stocks



Bayesian Learning

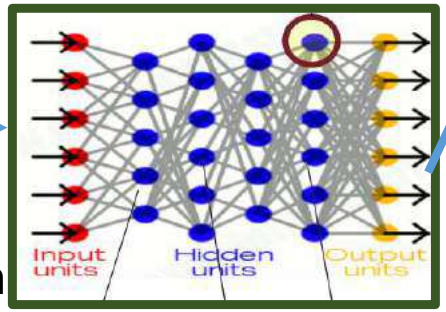


Report

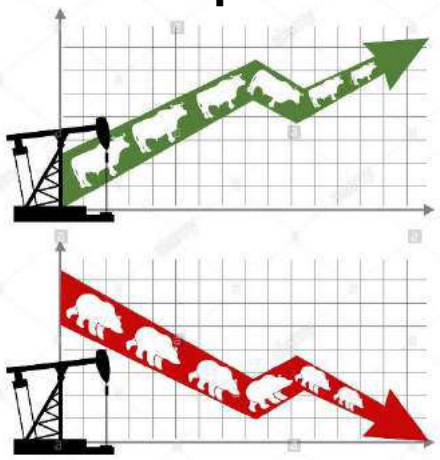


Annual Reports

Find Explanation



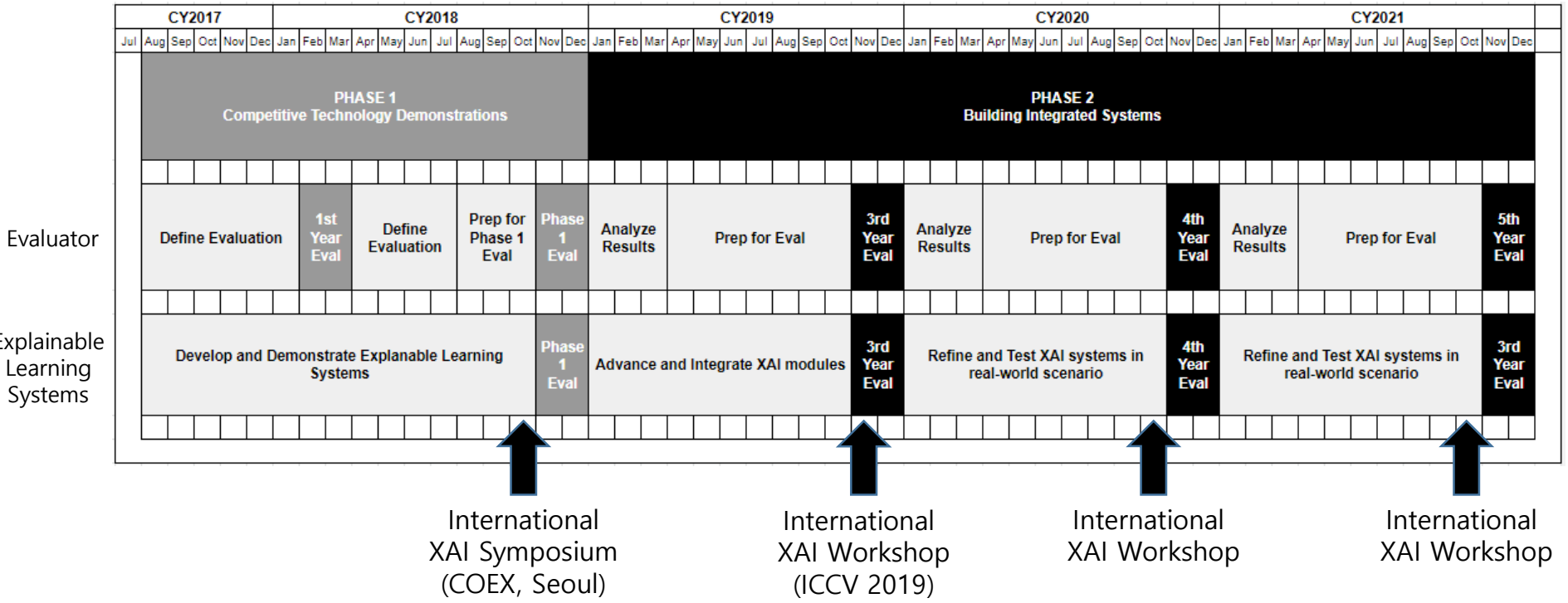
Deep Learning



Prediction

Read the Reports and Data and Explain It

Schedule


CY2017
CY2018
CY2019
CY2020
CY2021
Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
PHASE 1
Competitive Technology Demonstrations
PHASE 2
Building Integrated Systems
Evaluator
Define Evaluation
1st Year Eval
Define Evaluation
Prep for Phase 1 Eval
Phase 1 Eval
Analyze Results
Prep for Eval
3rd Year Eval
Analyze Results
Prep for Eval
4th Year Eval
Analyze Results
Prep for Eval
5th Year Eval
Explainable Learning Systems
Develop and Demonstrate Explanable Learning Systems
Phase 1 Eval
Advance and Integrate XAI modules
3rd Year Eval
Refine and Test XAI systems in real-world scenario
4th Year Eval
Refine and Test XAI systems in real-world scenario
3rd Year Eval
International XAI Symposium (COEX, Seoul)
International XAI Workshop (ICCV 2019)
International XAI Workshop
International XAI Workshop

International Meetings



2018 International XAI Symposium

Wish to Meet and Discuss!



2019 International XAI Workshop (will propose)

Publications – Journals and Conferences

1. Sungsoo Ahn, Michael Chertkov, Adrian Weller, and Jinwoo Shin, "*Bucket Renormalization for Approximate Inference*", **ICML** 2018
2. Hae Beom Lee, Eunho Yang, and Sung Ju Hwang, "*Deep Asymmetric Multi-task Feature Learning*", **ICML** 2018
3. 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
4. Wonseok Jeon, Seokin Seo, and Kee-Eung Kim "*A Bayesian Approach to Generative Adversarial Imitation Learning*", **NeurIPS** 2018
5. Jongmin Lee, Geon-hyeong Kim, Pascal Poupart, and Kee-Eung Kim "*Monte-Carlo Tree Search for Constrained POMDPs*", **NeurIPS** 2018
6. Jonghwan Mun, Kimin Lee, Jinwoo Shin, Bohyung Han, "*Learning to Specialize with Knowledge Distillation for Visual Question Answering*", **NeurIPS** 2018
7. Hajin Shim, Sung Ju Hwang and Eunho Yang, "*Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding*", **NeurIPS** 2018
8. Hae Beom Lee, Juho Lee, Saehoon Kim, Eunho Yang, and Sung Ju Hwang, "*DropMax : Adaptive Variational Softmax*", **NeurIPS** 2018
9. 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
10. Kimin Lee, Honglak Lee, Kibok Lee and Jinwoo Shin "*Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples*", **ICLR** 2018
11. 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

Publications in 2018 – Journals and Conferences

12. Saehoon Kim, Jungtaek Kim, and Seungjin Choi, "*On the Optimal Bit Complexity of Circulant Binary Embedding*", **AAAI** 2018
13. Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang, *Machine-Translated Knowledge Transfer for Commonsense Causal Reasoning*, **AAAI** 2018.
14. Reinald Kim Amplayo, Kyungjae Lee, Jinyeong Yeo, and Seung-won Hwang, "*Translations as Additional Contexts for Sentence Classification*", **IJCAI** 2018
15. Reinald Kim Amplayo, Jihyeok Kim, Sua Sung, and Seung-won Hwang, "*Cold-Start Aware User and Product Attention for Sentiment Classification*", **ACL** 2018
16. Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh, "*Subword-level Word Vector Representations for Korean*", **ACL** 2018
17. Reinald Kim Amplayo, Seonjae Lim, and Seung-won Hwang, "*Entity Commonsense Representation for Neural Abstractive Summarization*", **NAACL** 2018
18. JinYeong Bak and Alice Oh, "*Conversational Decision Making Model for Predicting King's Decision in the Annals of the Joseon Dynasty*", **EMNLP** 2018
19. Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, Seung-won Hwang, "*Visual Choice of Plausible Alternatives: An Evaluation of Image-based Commonsense Causal Reasoning*", **LREC** 2018
20. Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho, "*Attentive Semantic Alignment with Offset-Aware Correlation Kernels*", **ECCV** 2018
21. Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han, "*CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps*", **ECCV** 2018
22. Seong Tae Kim and Yong Man Ro, *"Facial Dynamics Interpreter Network: What are the Important Relations between Local Dynamics for Facial Trait Estimation?*", **ECCV** 2018

Publications in 2018 – Journals and Conferences

23. Paul Hongsuck Seo, Zhe Lin, Scott Cohen, Xiaohui Shen, and Bohyung Han, "*Progressive Attention Networks for Visual Attribute Prediction*", **BMVC** 2018
24. Jeany Son, Daniel Kim, Solae Lee, Suha Kwak, Minsu Cho, and Bohyung Han, "*Forget & Diversify: Regularized Refinement for Weakly Supervised Object Detection*", **ACCV** 2018
25. Kinam Kwon and HyunWook Park, "*A Learning-based Metal Artifacts Correction Method for MRI using Dual-Polarity Readout Gradients and Simulated Data*", **MICCAI** 2018
26. Sangmin S. Lee, Hak Gu Kim, Kihyun Kim, and Yong Man Ro, "*Adversarial spatial frequency domain critic learning for age and gender classification*", **ICIP** 2018
27. Seong Tae Kim, Jae-Hyeok Lee, Hakmin Lee, and Yong Man Ro, "*Visually Interpretable Deep Network for Diagnosis of Breast Masses on Mammograms*," **Physics in Medicine & Biology**, 2018
28. Eunsong Kang and Heung-il Seok, "*Probabilistic Source Separation on resting-state fMRI and Its Use for Early MCI Identification*", **MICCAI** 2018
29. Woo-Hyuck Song, Hong-Gyu Jung, In-Youb Gwak, and Seong-Whan Lee, "*Oblique Aerial Image Matching based on Iterative Simulation and Homography Evaluation*", **Pattern Recognition** 2018

Publications – Workshops and Demos

1. Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "*A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks*", **ICML 2018 workshop**
2. Jae-Hyeok Lee, Seong Tae Kim, Hakmin Lee, and Yong Man Ro, "*Feature2Mass: Visual Feature Processing in Latent Space for Realistic Labeled Mass Generation*", **ECCV 2018 workshop**
3. Anh Tong and Jaesik Choi, "*Discovering Relational Covariance Structures for Explaining Multiple Time Series*", **IJCAI 2018 workshop**
4. Subin Yi and Jaesik Choi, "*Learning Group Structure of Deep Neural Networks with an Expectation Maximization Method*", **ICDM 2018 workshop**
5. Minseop Park, Saehoon Kim, Jungtaek Kim, Yanbin Liu, and Seungjin Choi, "*TAEML: Task-adaptive ensemble of meta-learners*", **NeurIPS 2018 workshop**
6. Dong-Ok Won, Byung-Do Kim, Ho-Jung Kim, Tae-San Eom, Klaus-Robert Müller, and Seong-Whan Lee, "*Curly: An AI-based Curling Robot Successfully Competing in the Olympic Discipline of Curling*", **IJCAI 2018 Demo**

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

Thank you

jaesik@unist.ac.kr



설명가능인공지능:
사회를 위한 도구가 될 것인가?

고등과학원 초학제연구단
명지대학교 방목기초교육대학
전진권



"Black Box AI"

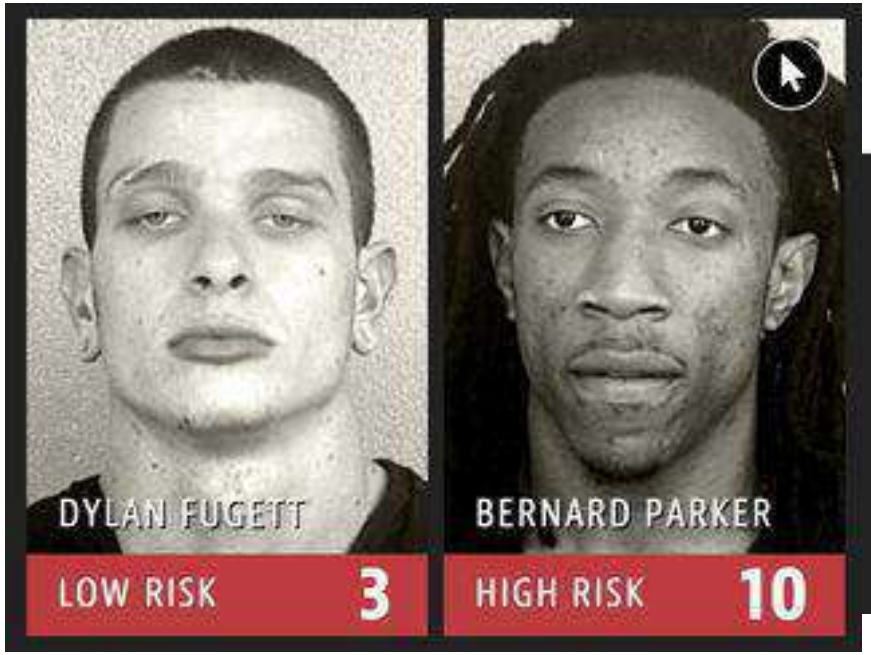
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

(Devlin et al., 2018)

Explainable or Interpretable AI (XAI)

Q: “If a machine learning model performs well, why not just trust the model and ignore why it made a certain decision?”

- The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks (Doshi-Velez and Kim 2017)
- AI that is explainable, provable, and transparent will be critical to establishing trust, meaning that for certain problems or tasks it is not enough to get the answer (the **what**). The model also has to give an explanation how it came to the answer (the **why**), because a correct prediction only partially solves your original problem (Molnar 2018).



COMPAS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

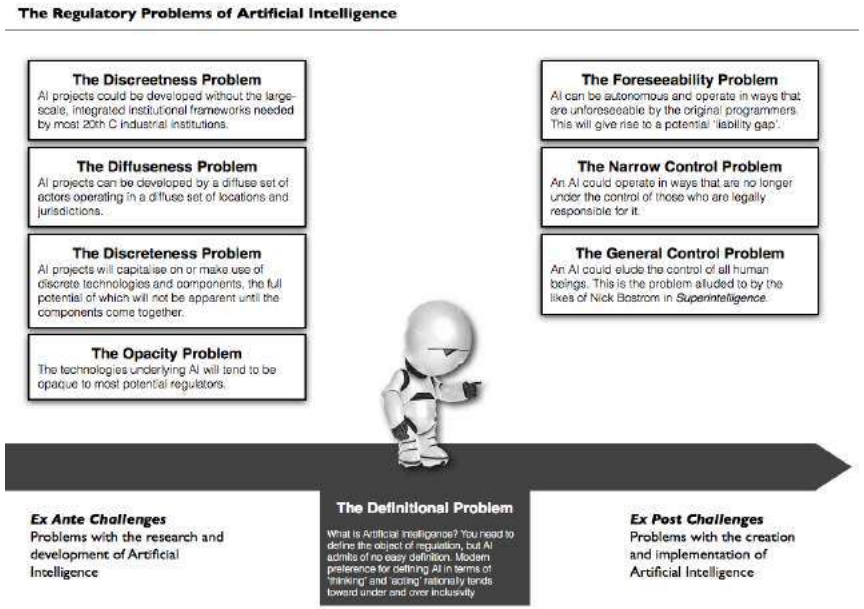
CNBC MARKETS BUSINESS NEWS INVESTING TECH POLITICS

Amazon scraps a secret A.I. recruiting tool that showed bias against women

- **Amazon.com's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.**

Algocracy

- Rule by Algorithm:
  "algorithm-based systems structure and constrain the opportunities for human participation in, and comprehension of, public decision-making" (Danaher 2016)



AI black boxes are scary when they:

1. make important decisions in areas that impact people personally – like in employment, lending, education, housing, etc.
2. are secretive, thus lacking in accountability
3. are destructive (Cathy O'Neill, 2016)

- "Core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (e.g "high stakes" domains) should no longer use "black box" AI and algorithmic systems (AI NOW, 2017)."

Explainable or Interpretable AI (XAI)

- XAI aims to help users understand, appropriately trust, and effectively manage the emerging generation of AI systems(DARPA, 2016).
- Interpretability as the ability to explain or to present in understandable terms to a human(Doshi-Velez & Kim, 2017).
- Interpretability is the degree to which a human can understand the cause of a decision (Miller 2017).

If we can understand how AI work:

- Fairness: Making sure the predictions are unbiased and not discriminating against protected groups (e.g. racial bias).
- Privacy: Ensuring that sensitive information in the data is protected.
- Reliability or Robustness: Test that small changes in the input don’t lead to big changes in the prediction.
- Trust: It is easier for humans to trust a system that explains its decisions compared to a black box.

(Doshi-Velez and Kim 2017)

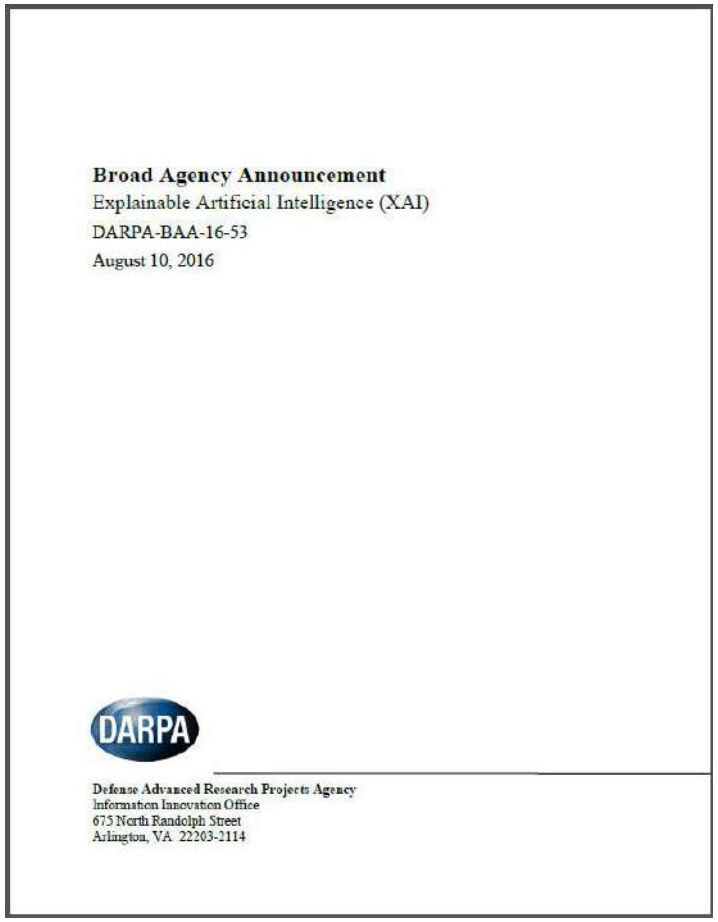
But, is XAI a panacea?

- Advocates talk too easily that XAI will help social problems.
- We think it is not a simple matter. There are still many challenges to be solved. And we need to be clear about the current technical situation and think about future tasks.

Outline

- The Current status of XAI
- Challenges and Limitations
- XAI and society

XAI is not new



Broad Agency Announcement
Explainable Artificial Intelligence (XAI)
DARPA-BAA-16-53
August 10, 2016

DARPA

Defense Advanced Research Projects Agency
Information Innovation Office
675 North Randolph Street
Arlington, VA 22203-2114

2016: DARPA's announcement of Funding Opportunity "Explainable Artificial Intelligence" (XAI)
POC: David Gunning

The concept of XAI is not new:
Fagan et al. 1980: computer-based consultation program (MYCIN)
- accurate
- manipulating a set of symbolic concepts
- interacting with natural language
- explaining the reasoning process
- funded by NIH

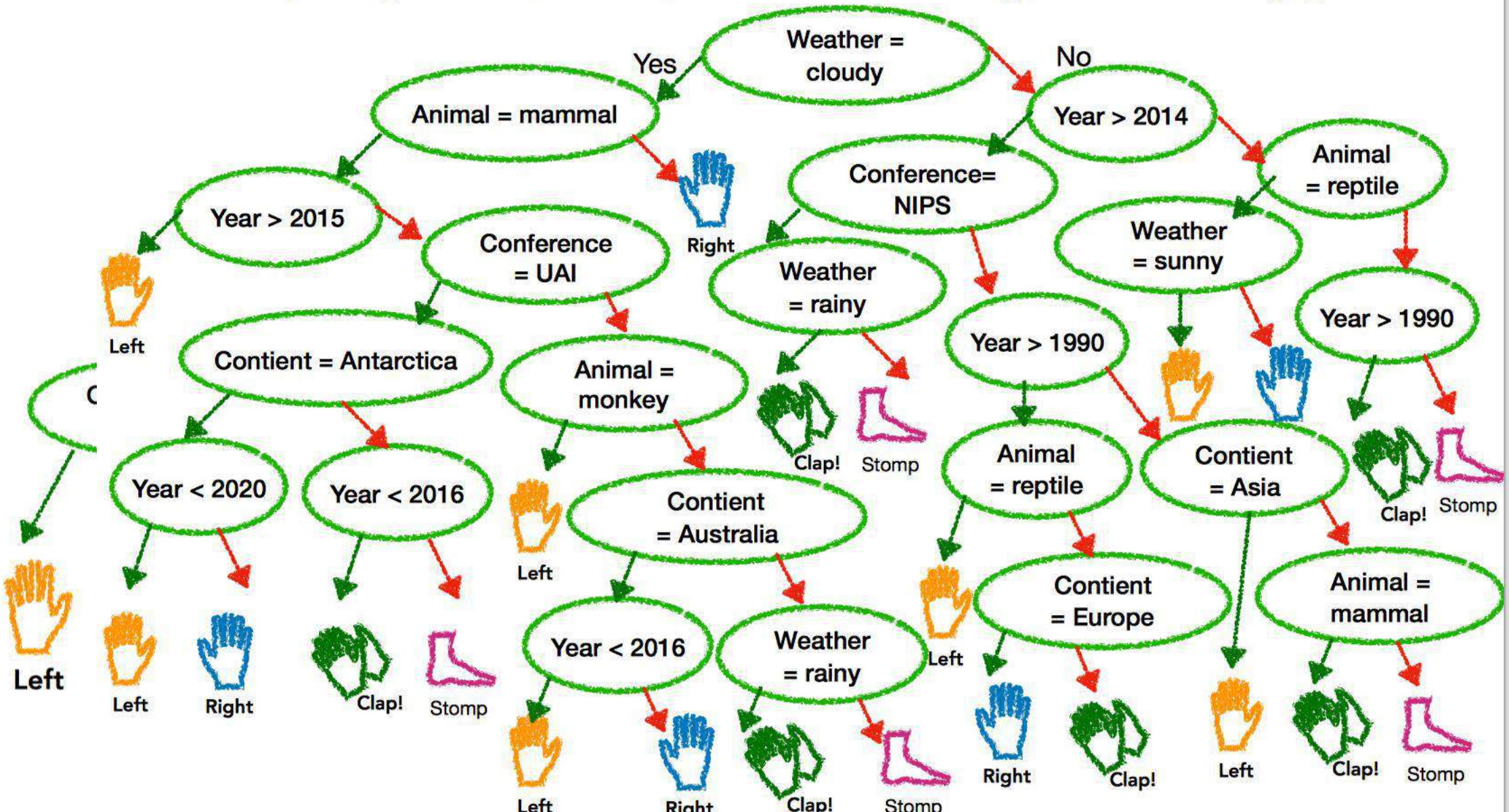
Swartout et al. 1991; Chandrasekaran and Swartout 1991
- "Explainable Expert System" (EES)
- funded by DARPA

Fisher and Mancuso 2004: explainable artificial intelligence
- military modelling; funded by Army

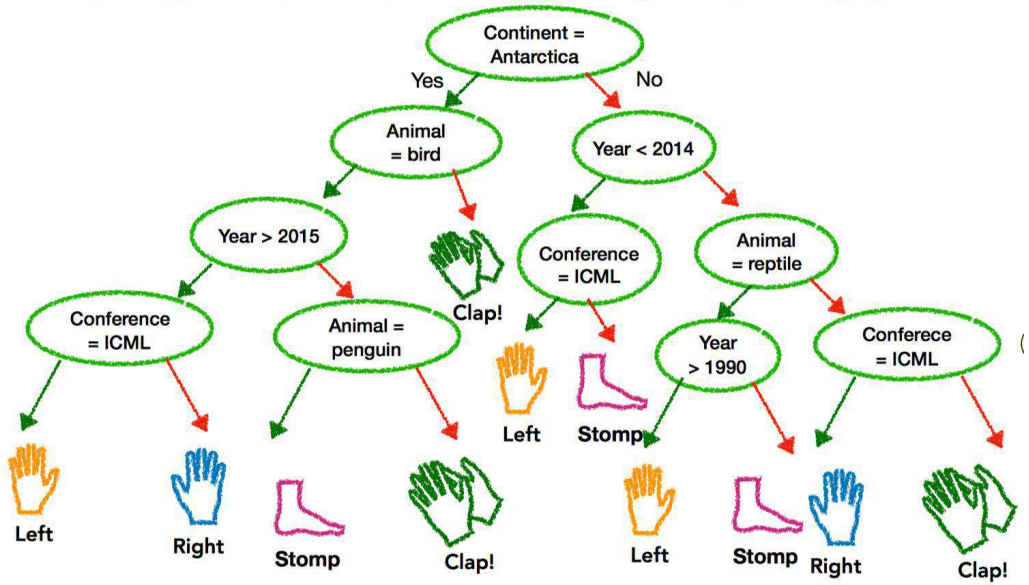
But, the current AI (such as Deep Learning) brings new challenges

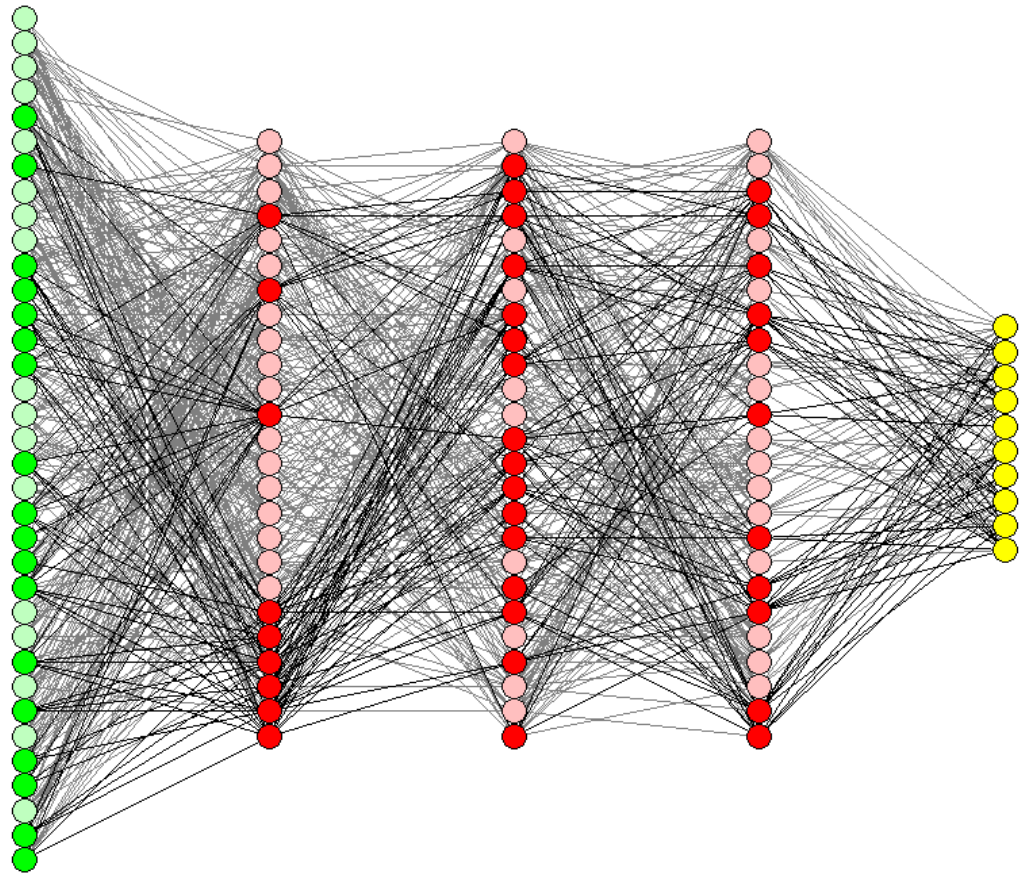
- In the last century, the most obvious approach to software transparency was the disclosure of the source code.
- But, “Machine learning, one increasingly popular approach to automated decision-making, is particularly ill-suited to source code analysis because it involves situations where the decisional rule itself emerges automatically from the specific data under analysis” (Kroll 2016)
- AI is too complex for humans to understand

Input: [ICML, 2017, Australia, Kangaroo, Sunny]



Input: [ICML, 2017, Australia, Kangaroo, Sunny]



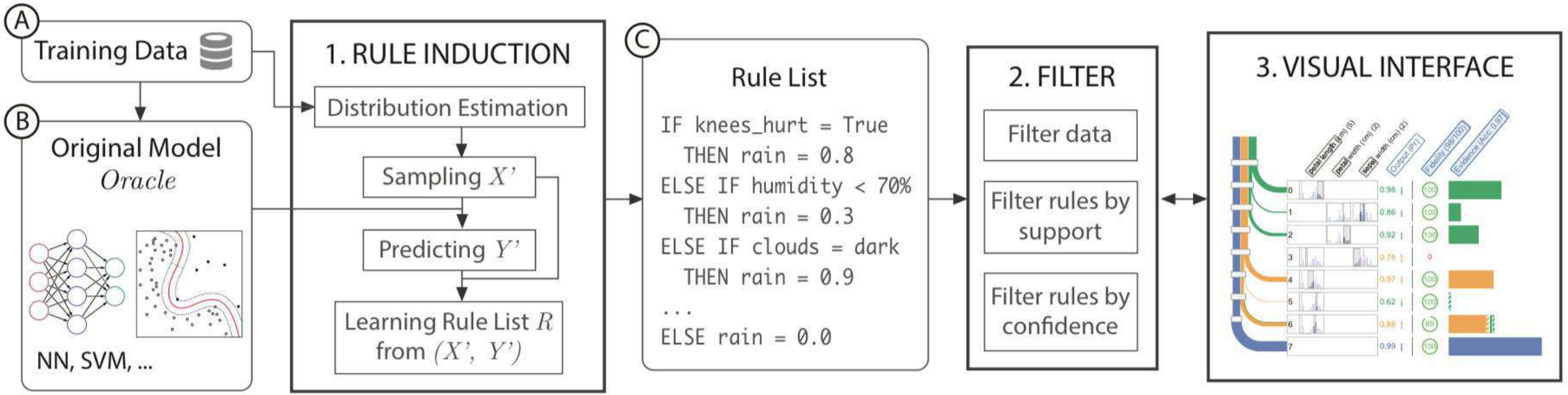


Rule-extraction

- If-then rule: IF $X \epsilon S(i)$ THEN $Y = y(i)$
- M-of-N rule: IF M of {N} THEN Z

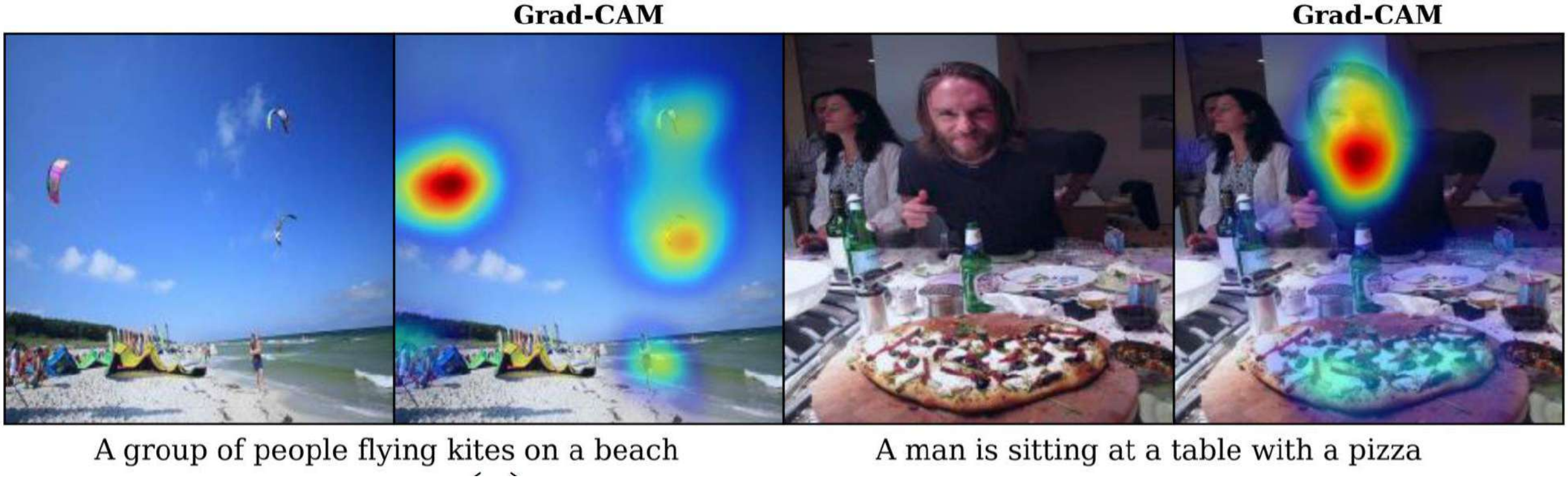
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

- Decomposition method: CRED, FERNN, KT, DeepRED
- Pedagogical method: RxREN, TREPAN, ANN-DT



Ming et al. 2018

Attribution method: Saliency Map



Selvaraju et al., 2017



Question : What vegetable is on the plate ?
Predicted Answer : broccoli

Question : What color is the plate ?
Predicted Answer : white

Question : Is there meat in this dish ?
Predicted Answer : no



Question : Where is the player ?
Predicted Answer : tennis court

Question : What does the man wear on his arms ?
Predicted Answer : tennis racket

Question : What sport is this ?
Predicted Answer : tennis

Goyal et al., 2016

Intrinsic method

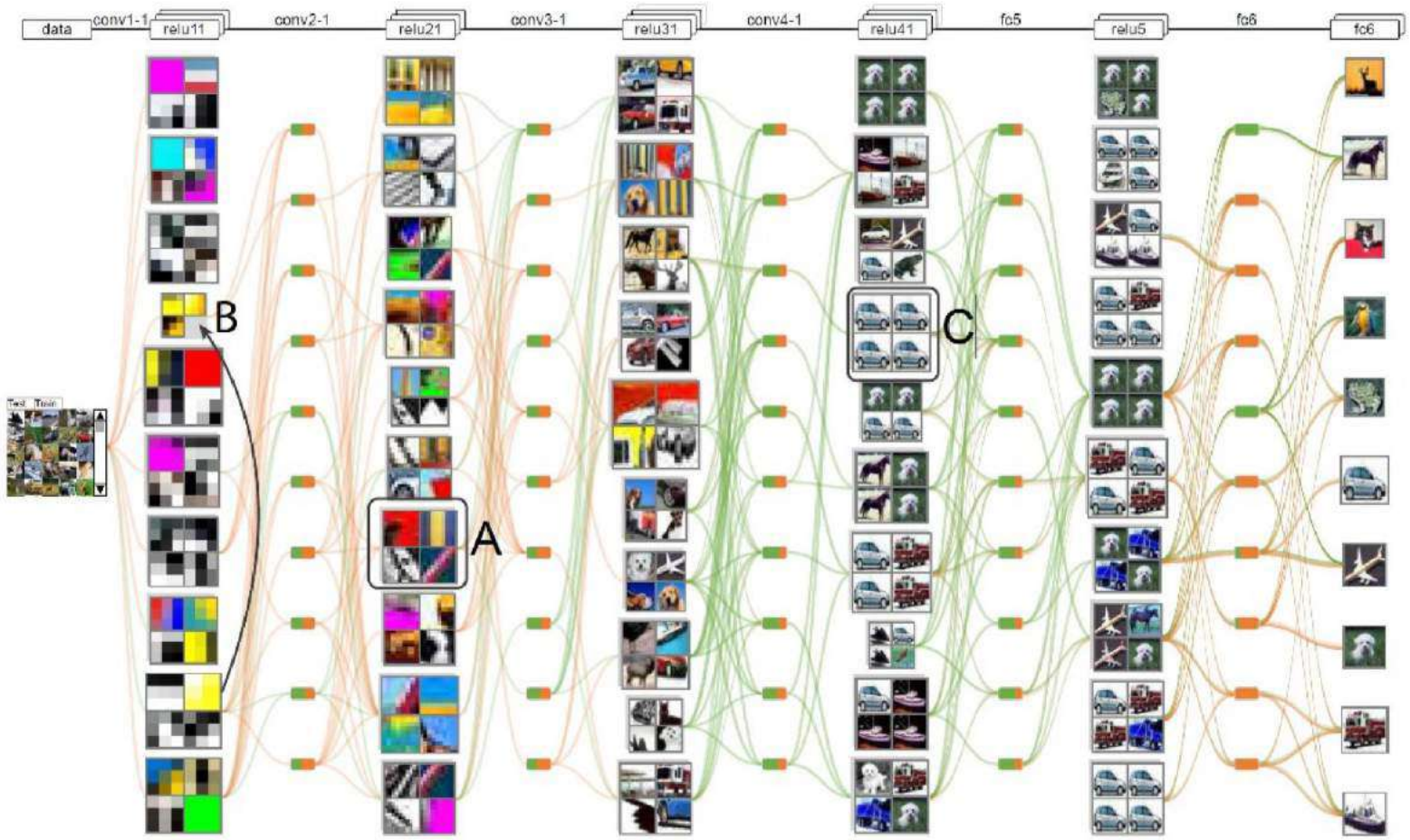
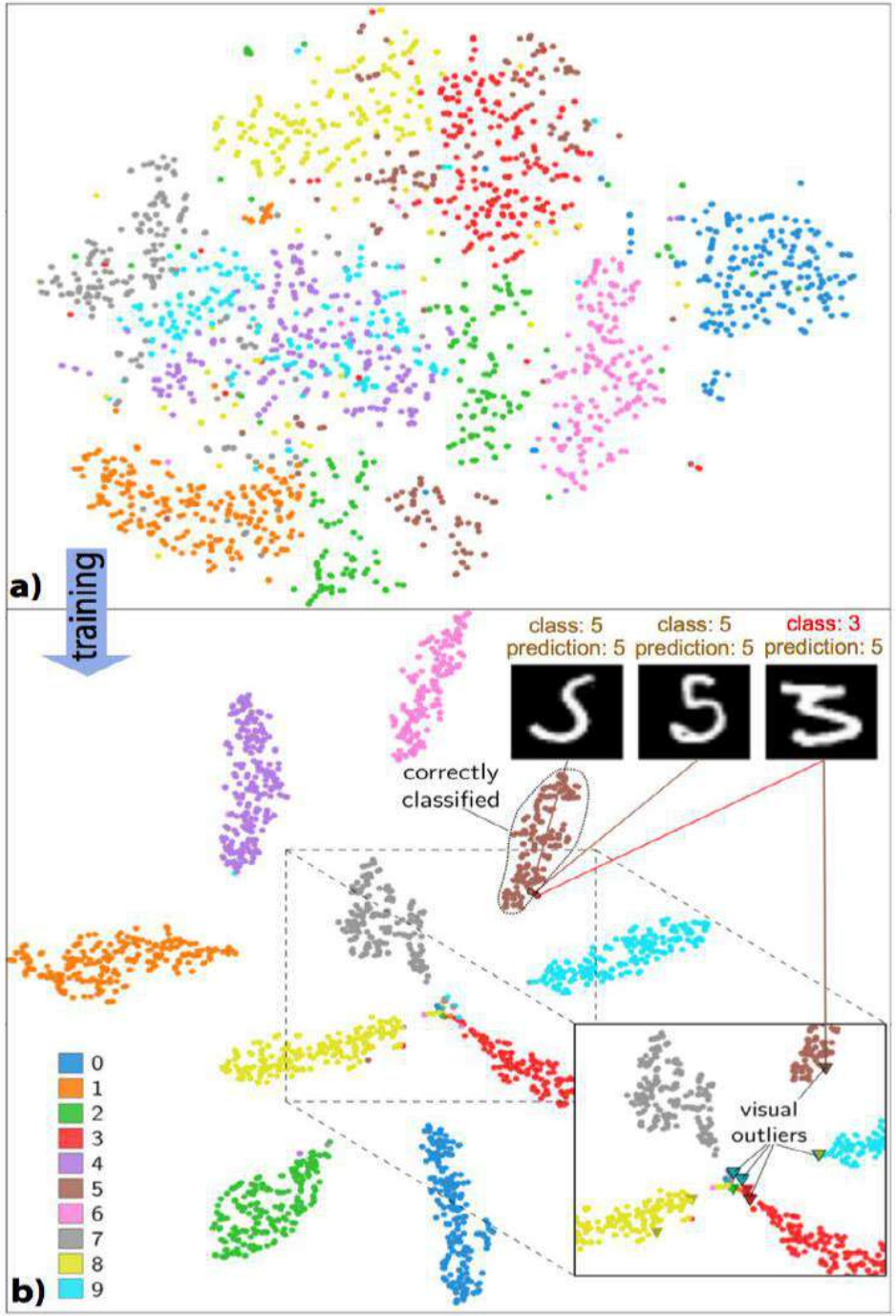


Figure 3. Overview of CNNVis [8]. The system employs a DAG to visually illustrate the inner workings of a CNN training process. It also performs bi-clustering to visualize the strength of interactions between filters of adjacent layers, along with their most activated images.

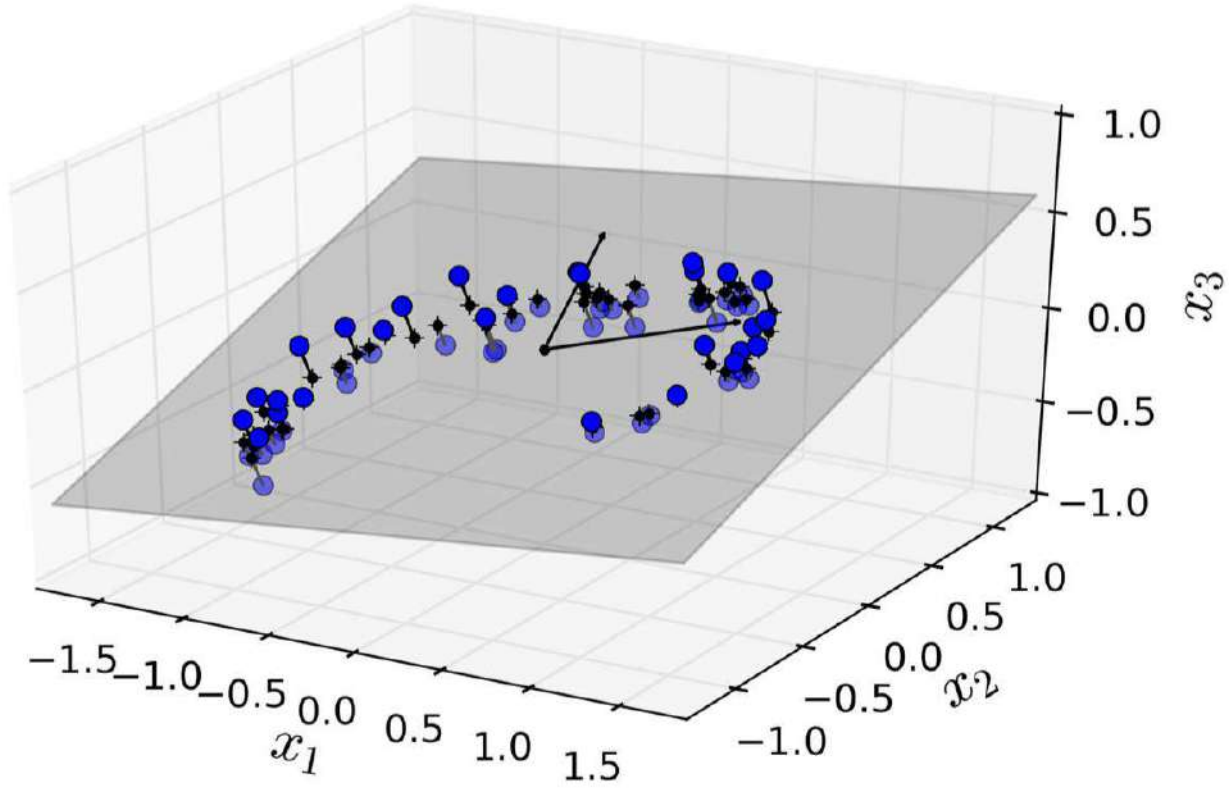
CNNvis. Choo & Liu, 2018



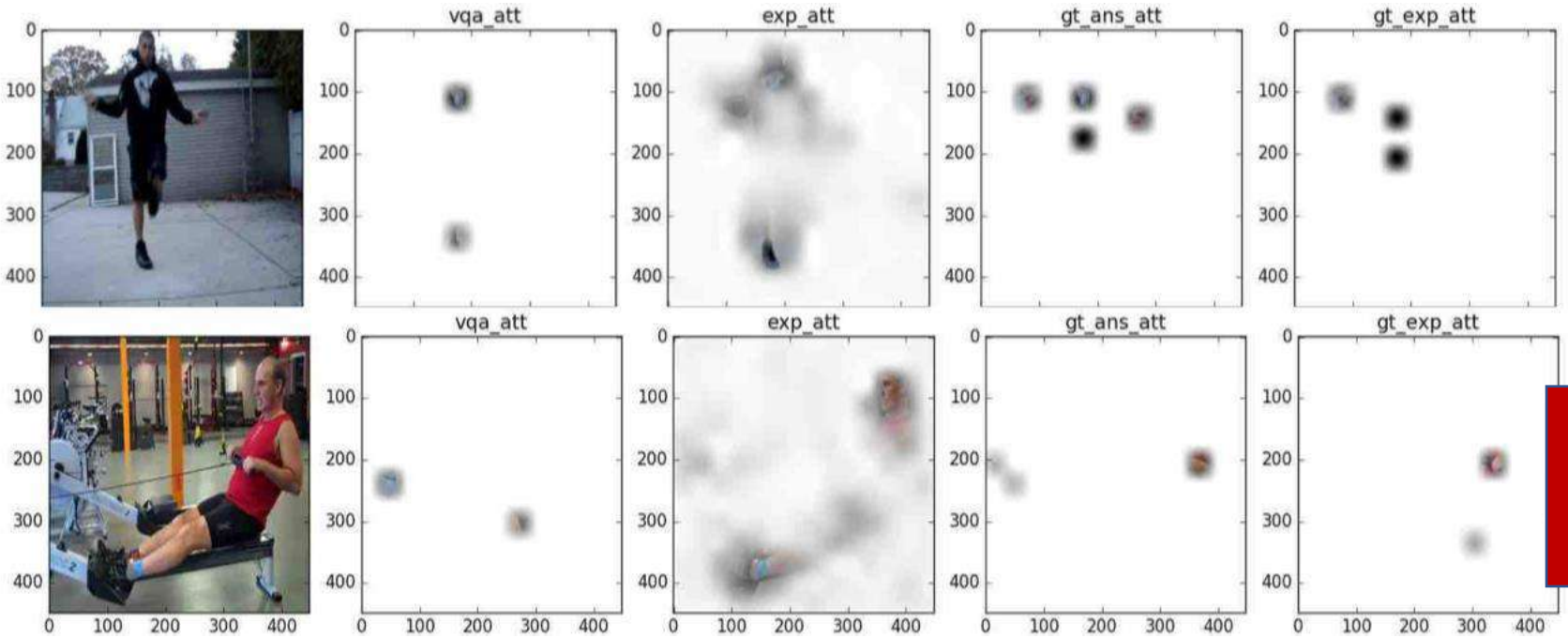
Rauber et al., 2017

Roughly, XAI is a model for a model

- In principle, all information about processes of AI is knowledgeable (e.g., log analysis). The problem is this information is vast and dynamic.
- XAI is a model about this information. A Typical XAI picks a feature, and compresses the information into the human understandable form
- This is not a true description but a useful approximation



Reason of the decision



Rope skipping

Because he is jumping up and down while swinging a rope over his head

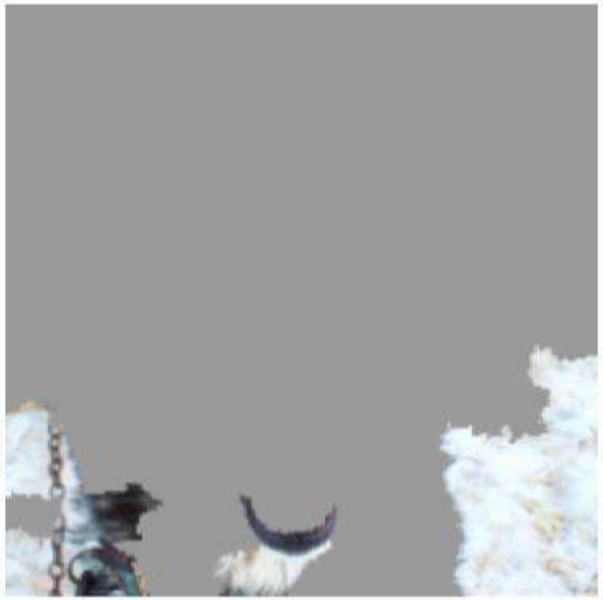
Rowing, stationary

Because he is sitting on a rowing machine and leaning back on the handles.

Verification of machine learning



(a) Husky classified as wolf



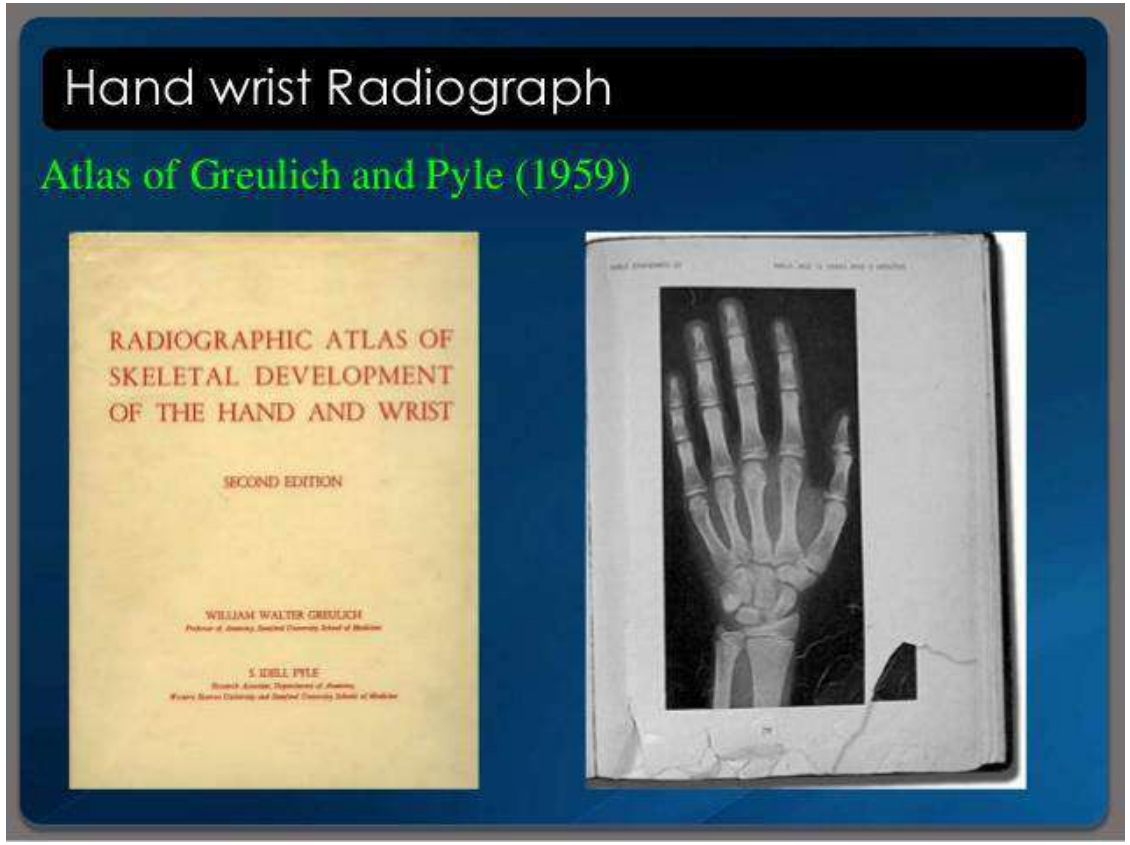
(b) Explanation

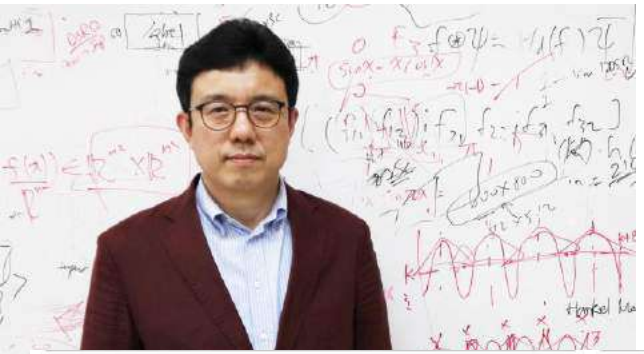
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

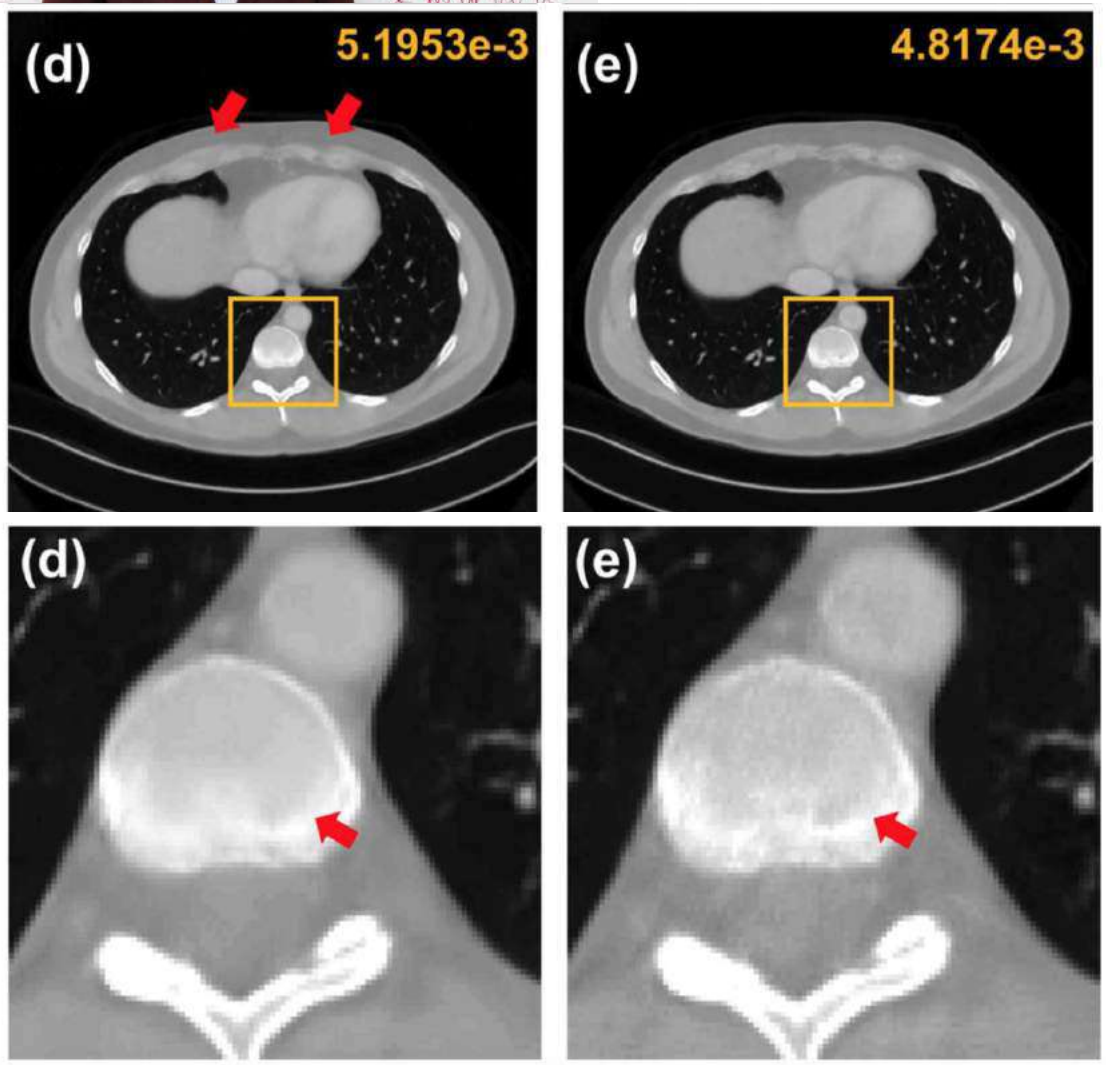
LIME. Ribeiro et al., 2016

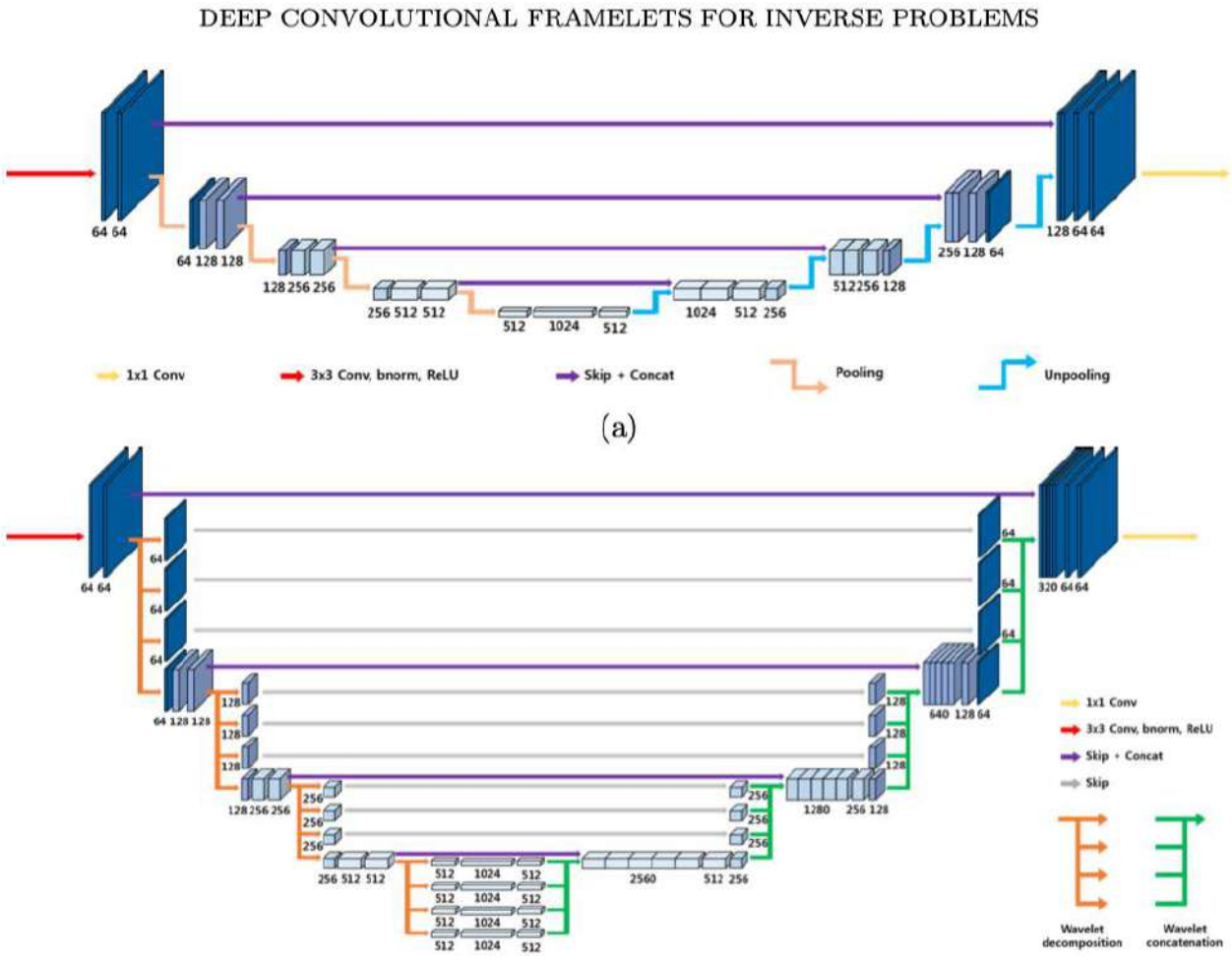


Vuno, <bone age>






(d)
5.1953e-3
(e)
4.8174e-3
(d)
(e)

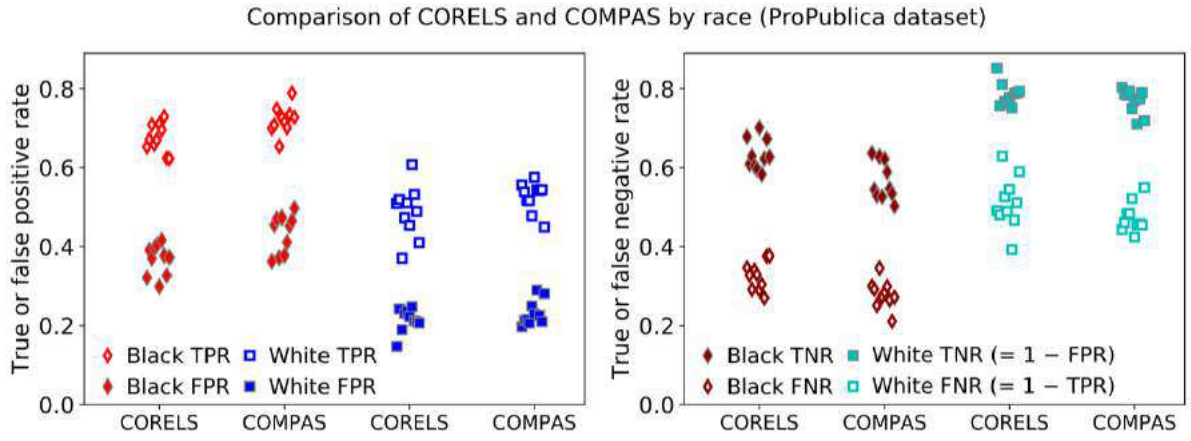

DEEP CONVOLUTIONAL FRAMELETS FOR INVERSE PROBLEMS
64 64
64 128 128
128 256 256
256 512 512
512
1024
512
1024 512 256
512256 128
256 128 64
128 64 64
1x1 Conv
3x3 Conv, bnorm, ReLU
Skip + Concat
Pooling
Unpooling
(a)
64 64
64 128 128
128 256 256
256 512 512
512 1024 512
512 1024 512
512 1024 512
2560
512 256
1280 256 128
640 128 64
320 64 64
1x1 Conv
3x3 Conv, bnorm, ReLU
Skip + Concat
Skip
Wavelet decomposition
Wavelet concatenation

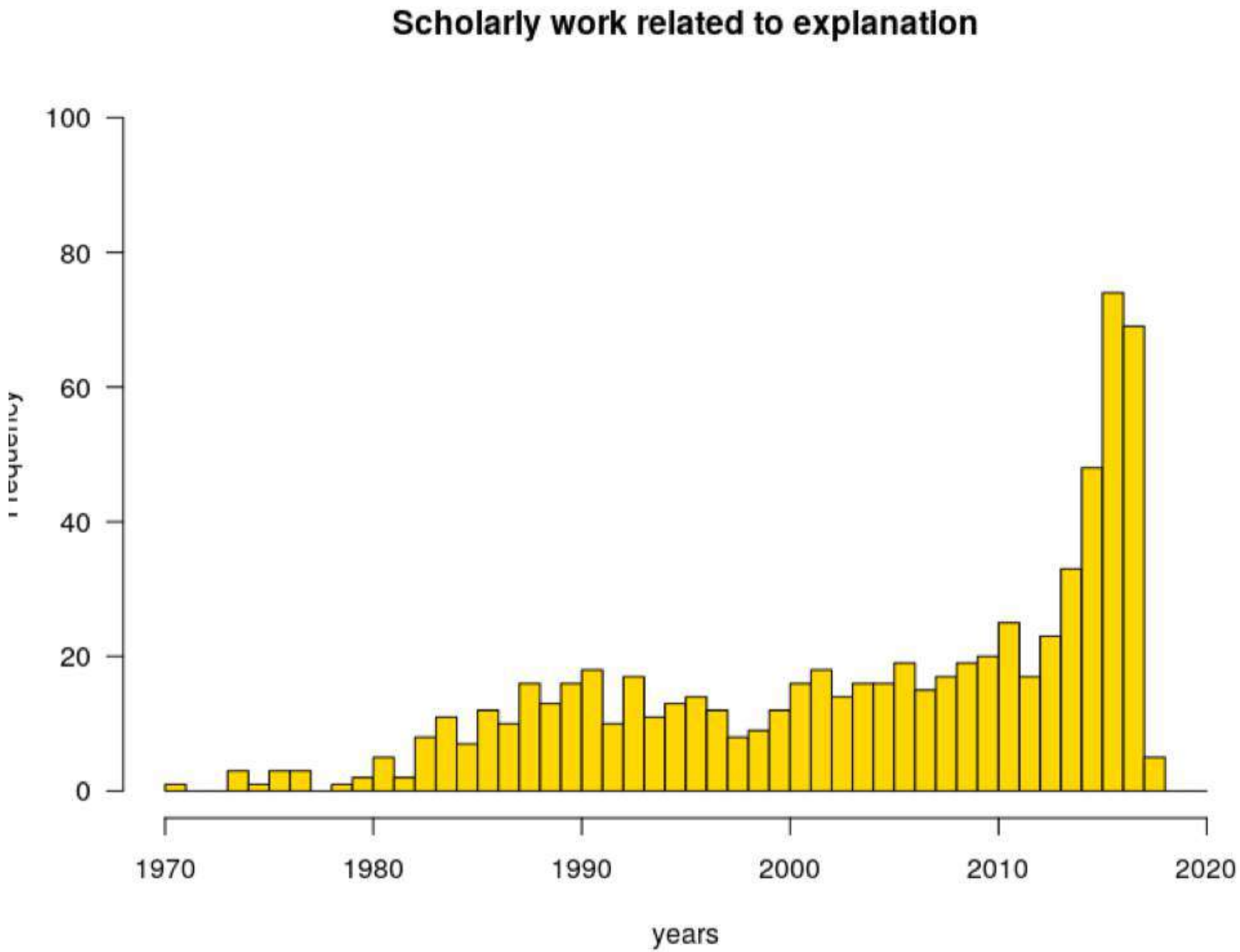
(Ye et al., 2018)

XAI vs. COMPAS

- CORELS (Angelino et al. 2018)
- A rule list model:

if $(age = 21 - 22)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no




Scholarly work related to explanation
100
80
60
40
20
0
1970
1980
1990
2000
2010
2020
years

What is the “explanation”?

- Most of XAI aim the accurate description of input-output relation.
- e.g.) Feature selection: There are infinite feature for representing AI process. Which is the best indicator of the whole process?
- A good XAI model: Accurate prediction of the result
- Good explanations are truthful: There is evidence that accuracy is also considered an important criterion of evaluation by the lay people (Yin et al. 2018).
- However, is it enough?

However, an explanation is more than prediction.

- Explanation is different in lay people's mind
- "Many artificial intelligence and machine learning publications and methods claim to be about 'interpretable' or 'explainable' methods, yet often this claim is only based on the authors intuition instead of hard facts and research(Miller 2017)."
- Social psychology has studied enormous researches on human understanding. However, they are not applied enough to XAI.

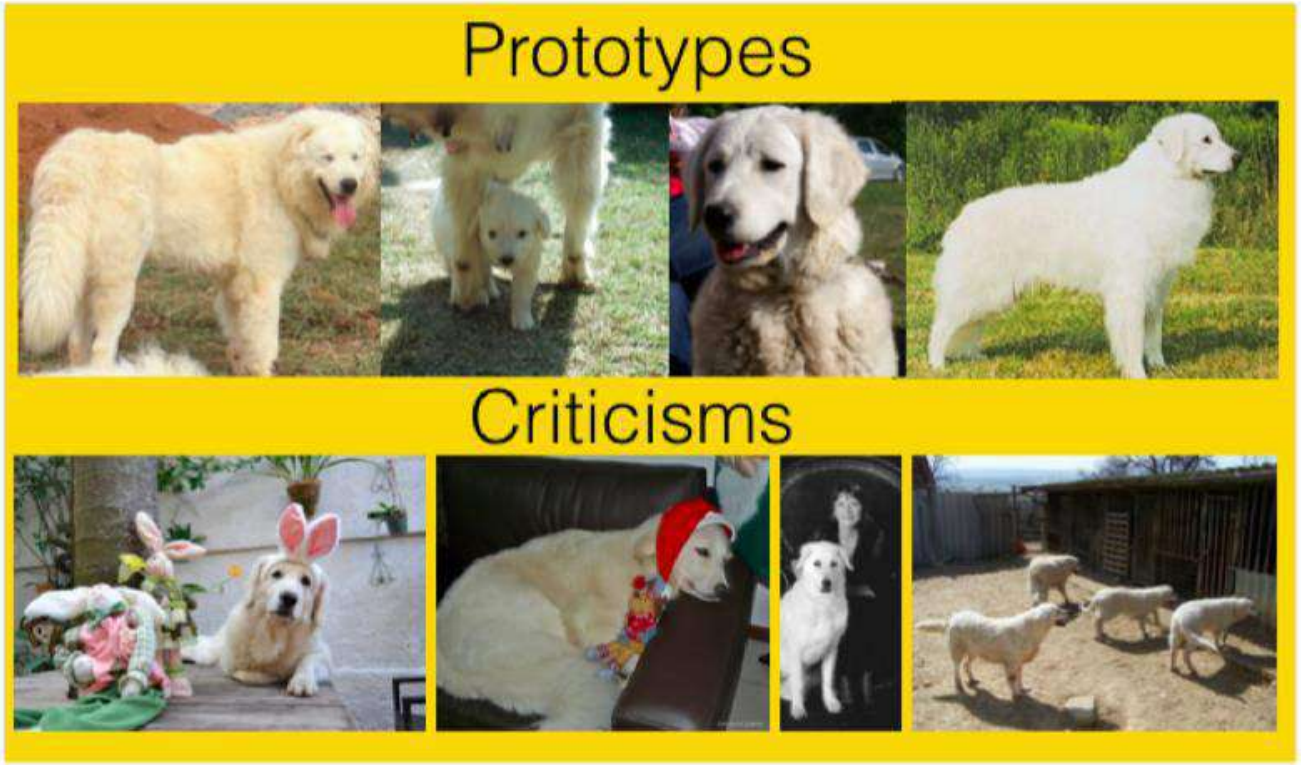
Good Explanations to Humans

- Explanations are selected: People don’t expect explanations to cover the actual and complete list of causes of an event. We are used to selecting one or two causes from a huge number of possible causes as THE explanation. Make the explanation very short, give only 1 to 3 reasons, even if the world is more complex.
- Generally, there is a trade-off between accuracy and interpretability. e.g., Accuracy requires more complex model, more features, more causal information, while Interpretability requires simpler model, more sparsity, one or two of causal information. It is hard to satisfy both.
- Explanations focus on the abnormal: People focus more on abnormal causes to explain events (Kahnemann 1981), even if other ‘normal’ features have the same influence on the prediction as the abnormal one. Good explanations are coherent with prior beliefs of the explainee

Explanations are contrastive

- contrastive explanation (Lipton 1990)
  - why P? ➔ Why P (fact), rather than Q (foil)?
- “Why did Elizabeth open the door?
  - “Why did Elizabeth open the door, rather than leave it closed?”
  - “Why did Elizabeth open the door rather than the window?”
  - “Why did Elizabeth open the door rather than Michael opening it? ”
- Contrastive explanations are easier to understand than complete explanations. Researches try to develop contrastive XAI (e.g. MMD-critic, Kim et al. 2016). However, it is still just algorithm dependent.
- Right contrastive XAI needs a proper reference set, which reflects people's real interests.

MMC-critic







Been Kim
Google Brain

Subjects performed the best using Proto and Criticism Condition (M=87.5%, SD=20%). The performance with Proto Only Condition was relatively similar (M=75%, SD=41%), while that with

(Kim et al., 2016)

Humans in XAI research

- Recently, researchers have introduced human experimental studies in XAI (Ribeiro et al., 2016; Kim et al. 2018). However, here, human beings are passive responders to given scenarios, but are not a subject actively engaging the situation.
- The XAI of this approach can take "power" away from the public. Here, XAI is merely a device in which a company or an organization persuades a user.
- XAI should be a "disclosure devices" (Hansen and Flyverbom 2015), not a "persuasion device".

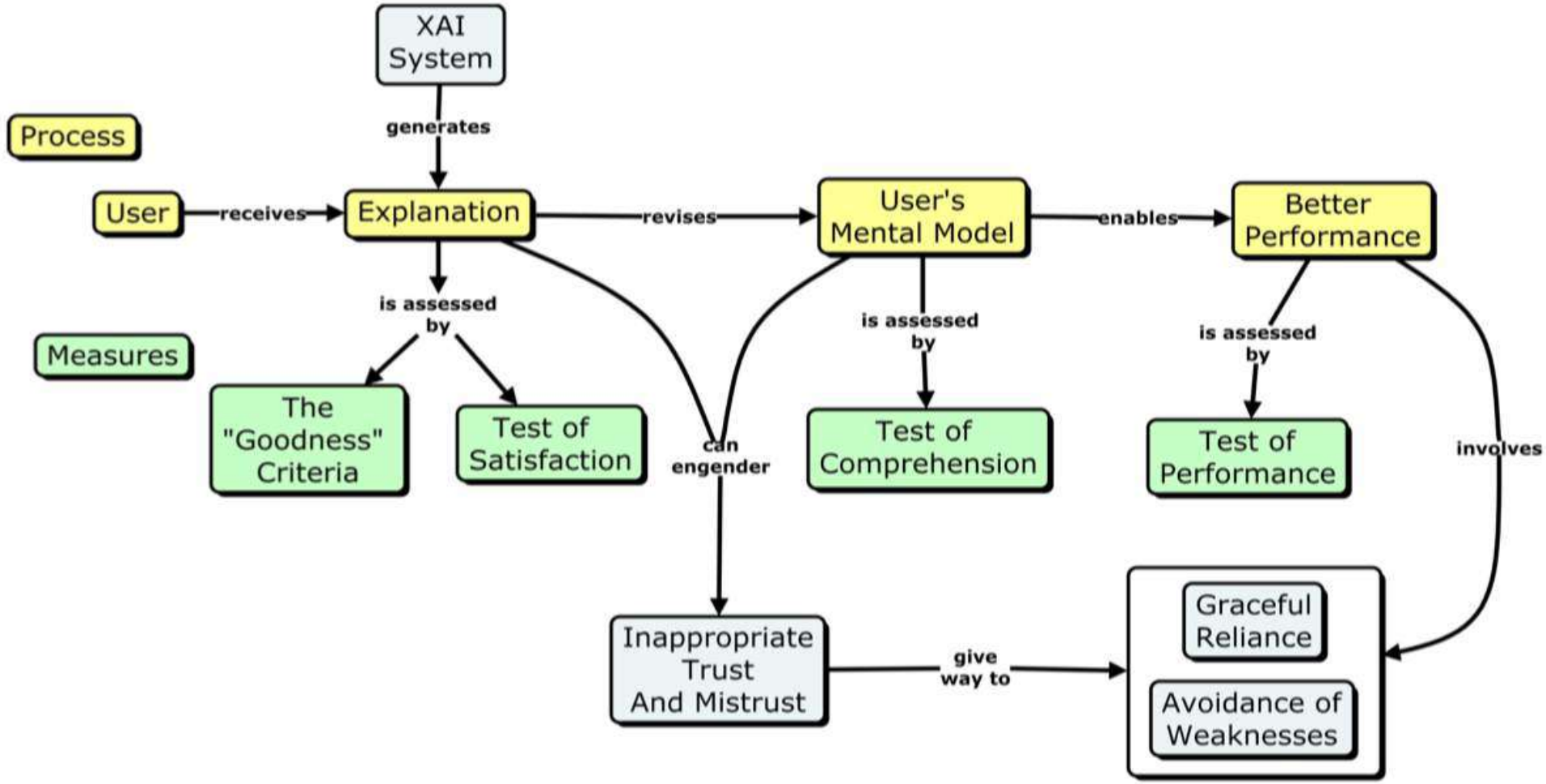


Figure 8.1. DARPA's XAI Evaluation Framework. Explanations should induce better mental models and performance, which produce appropriate trust in the system.

Concerns arise over manipulation

- XAI can change people's trust level, without changing the results(“nudged to take a preferred action”, Mittelstadt 2018).
- What if Google only provides a XAI depends on its favorable features? Who determines the criteria is a important issue for fairness. It can not be determined by algorithms alone.





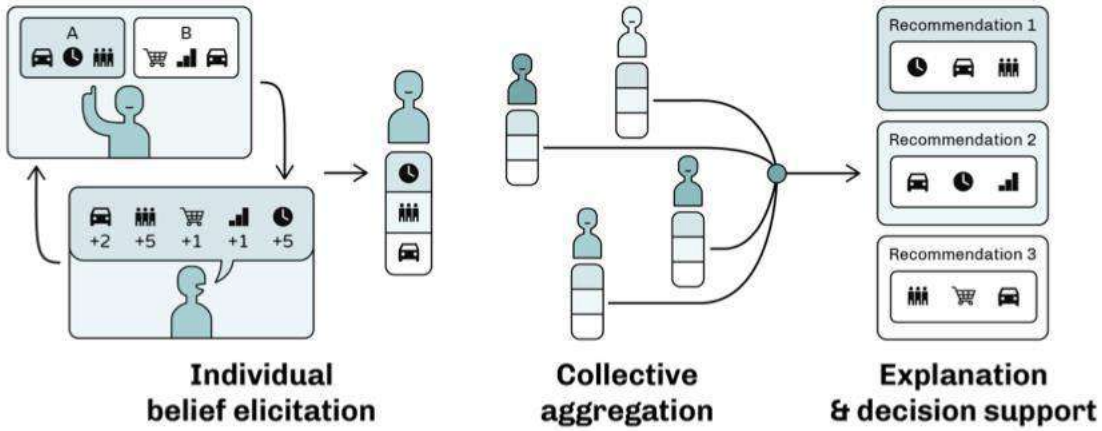
Good explanations are social

- In the social setting, people's interest is not just an explanation, but fairness. For example, if a loan was rejected by an automated algorithm, then the explanation that the loan was rejected due to lack of collateral is not sufficient. The real question is whether the loan screening was conducted fairly or not.
- "An algorithmic system is not just code and data but an assemblage of human and nonhuman actors—of institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally observable, semiautonomous action" (Ananny, 2015: 93)

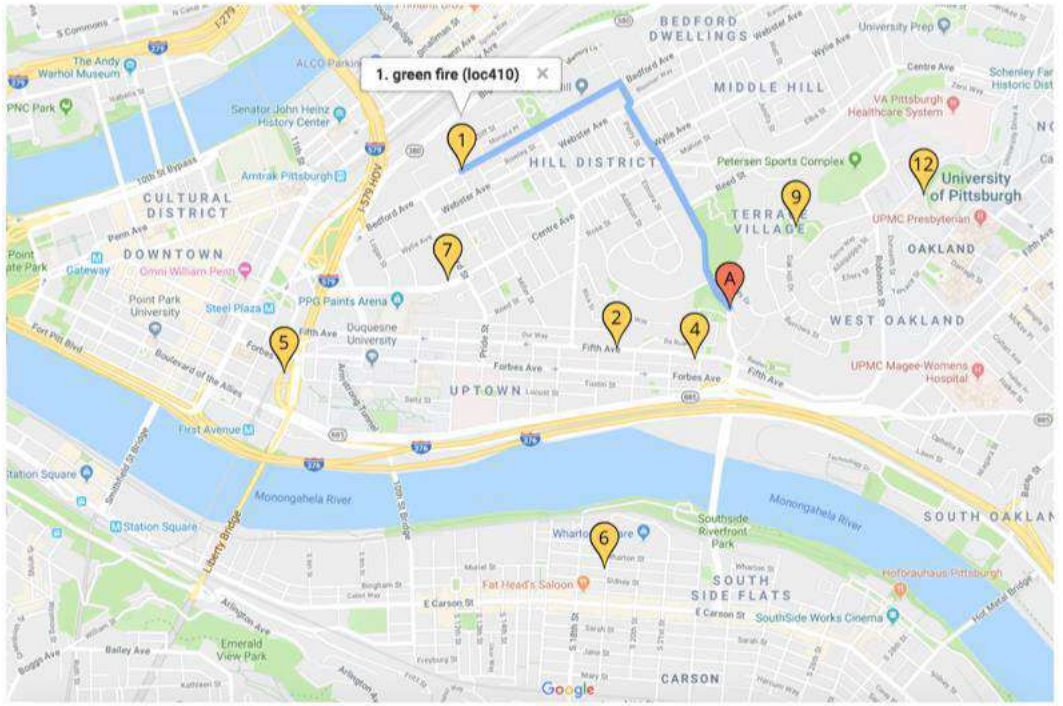
Before designing the XAI

- Who is the stakeholders?
- Which value should be prioritized?
- How should gains and losses be distributed?
- TRADE-OFF between Explanatory Power and Accuracy
- Understand the Explanation must be Local, Cultural, Context-dependent, Compromise of Values.

Human in the Loop (human/user-centered AI, Interactive machine learning)



WeBuildAI (Lee et al., 2018)



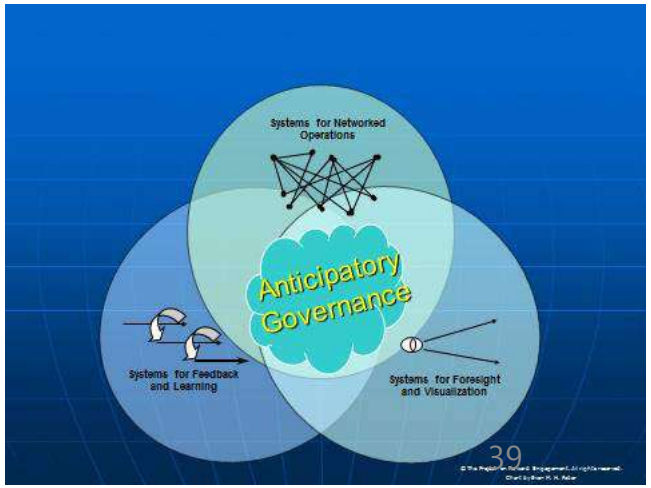
Anticipatory governance (Guston, 2014)

Foresight: Unlike prediction, foresight uses various means to anticipate implications of diverse plausible futures

Engagement: Public debate and deliberation aimed at both understanding and informing policy decisions

Integration: Critical reflection by experts that expands the scope of their decision frames and alternatives

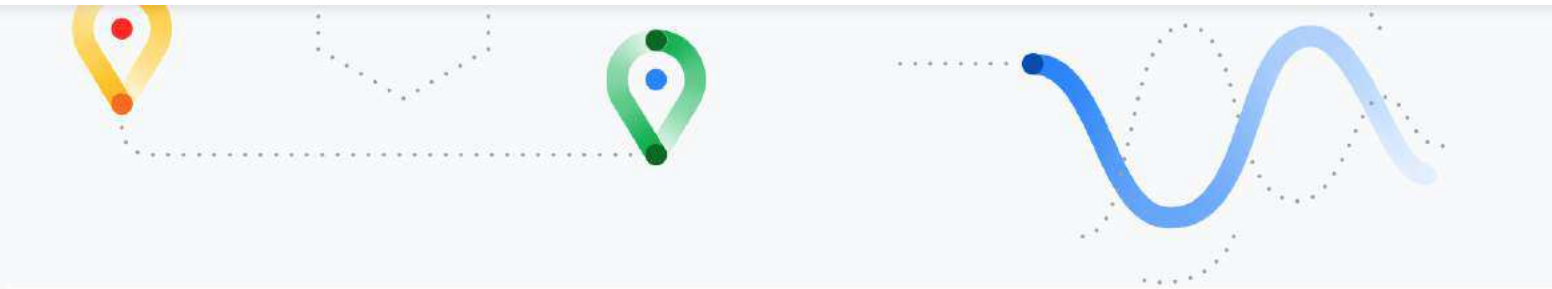
Ensemble-ization



Why do we need this effort?

- In digital contexts, transparency is not simply about revealing information or keeping secrets but continually deploying, configuring, and resisting platforms, algorithms, and machine learning protocols that manage visibility (Crawford 2016).
- An understandable AI enables people to be a part of the decision-making process in an AI-driven enterprise. Also critical to the Understandable AI process is the integration of non-data scientists to the development and design of AI products (Chowdhury 2018)





RESEARCH › TEAMS & FOCUS AREAS › BRAIN TEAM ›

arch (PAIR)

HOME ABOUT

n to make AI

nd fair.

xaic

Bringing to
and practit
accountabi

/ Project

What is XAI?

The goal is to create a suite of new or modified machine learning techniques that produce explainable models that, when combined with effective explanation techniques, enable and users to understand, appropriately trust, and effectively manage the emerging generation of Artificial Intelligence(AI) systems.

Our project consists of 4 parts.

Concluding remarks

- XAI aims to make AI “a glass box” and is attracting great attention.
- However, unlike expectations, it does not immediately bring governance improvements: There is still a lack of consideration for humans, especially here humans are passive subjects who do not actively involved in technology. It can harm the health of civil society.
- Watch out Algorithm-Centrism: The algorithm can not be the answer to all problems. A conversation (or argumentation) with civil society or other disciplines is still critical for “XAI for governance”
- For XAI to be applied to society: a more human-friendly and participatory XAI is needed.

Concluding remarks

Distinguish “Interpretable” and “Explainable”:

“We take the stance that interpretability alone is insufficient. In order for humans to trust black-box methods, we need explainability – models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. While interpretability is a substantial first step, these mechanisms need to also be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited.”(Gilpin et al., 2018)