国立研究開発法人
科学技術振興機構
Japan Science and Technology Agency

機械学習の説明可能性への取り組み

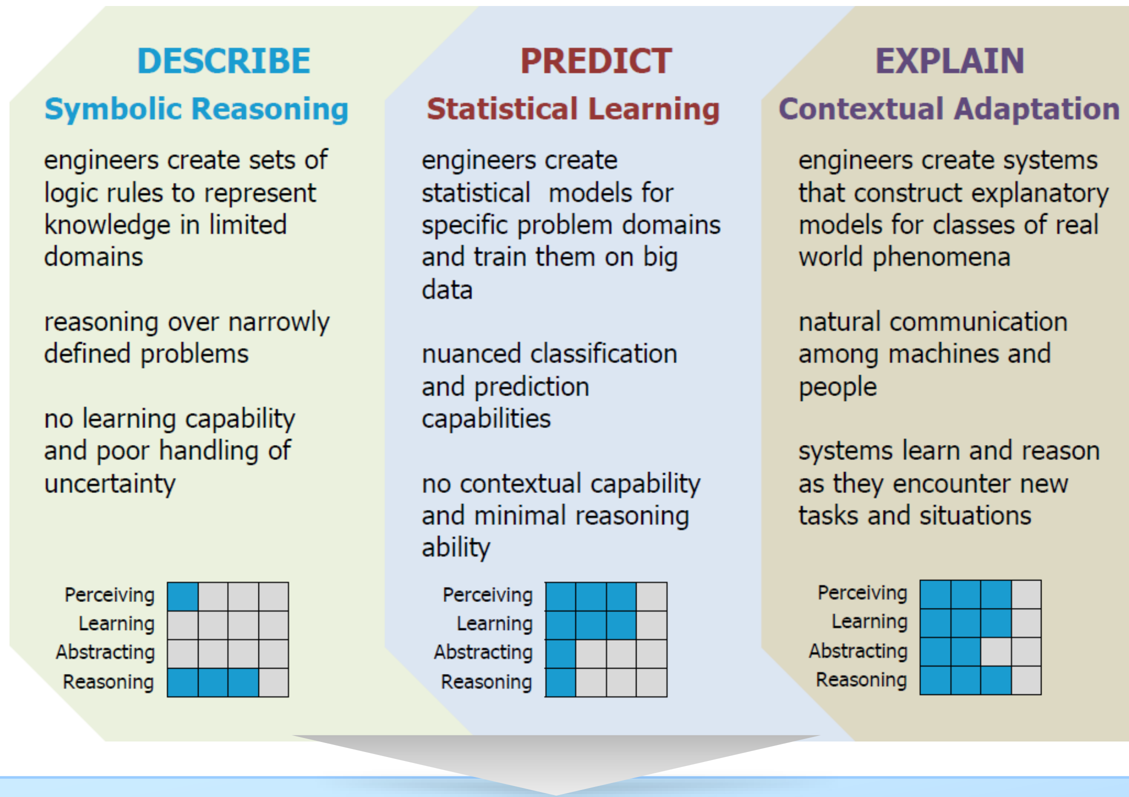
ー DARPA XAI プロジェクトを中心に ー

本発表では，報告者の私見を交えながら DARPA XAIプロジェクトを紹介します
進行中のPJであるため，情報が部分的かつ不正確である点について予めご承知おき下さい

川村 隆浩
国立研究開発法人 科学技術振興機構 特任フェロー

※本資料における図表はhttps://www.darpa.mil/program/explainable-artificial-intelligence より抜粋

1. XAIプロジェクト概要



現在をAI第3の波Contextual Adaptationと位置付け，今後現れるであろうAI，特に機械学習に基づくパートナーを*warfighters*が理解し，適切に信頼し，効率的に管理することを目指す．

そのため，高度な学習機能を維持しながら，より説明可能なモデルを生成する機械学習技術を開発，同時に，最新のHCI技術によってモデルをエンドユーザーが理解可能で有用な説明に翻訳する．

⇒精度と説明可能性にはトレードオフがあることを陽に述べており，両方を評価するとしている．

⇒インタフェースとの統合，およびその専門家との連携が初めから企図されている．

1.1 タスク設定

学習対象　　　　　　　　　　　　説明の使い方

	タスク実行モデルの学習	認識結果やアクションをユーザに説明	タスク実行にあたって説明を参照	

タスク1

Data Analytics

分類問題

マルチメディアデータ

説明可能モデル → 説明インタフェース ⇄ 推薦 / 説明

Images

分析官は，大量のマルチメディアデータの中から特定の物体を探している

大規模データセット内の特定の物体を認識

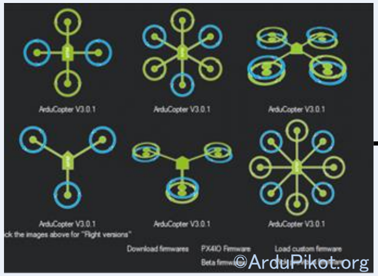
何故，推薦したか（しなかったか）を説明

分析官は，報告・追跡する物体を決定

タスク2

Autonomy

強化学習問題



ArduPilot & SITLシミュレーション

説明可能モデル → 説明インタフェース ⇄ アクション / 説明

オペレータは、一連のミッションを達成するために自律システムを指揮している

シミュレートされたミッションにおいて決定方針を学習

実行後にアクションを説明

操作者は，今後どのタスクを委譲するか決定

1.1 タスク設定

2つのタスクは，2つの重要な機械学習のアプローチ（分類問題と強化学習※）に対応し，2つの重要なDARPAのミッション（機密情報分析と自律システム）に対応している．

- **Data Analytics**
  - **分類問題** AND **機密情報分析に対応**
  - **対象はマルチメディアデータ**
  - **説明の目的は，分析官がどのターゲットを選ぶかを決めるための材料提供**

    ⇒例えば，敵と判断した一理由としては銃の輪郭を強調表示する，などか

    NN**自体を説明しなくてはいけないという漠としたイメージよりもかなり現実的な設定**
- **Autonomy**
  - **強化学習** AND **自律システムに対応**
  - **対象はドローン，ロボットなどの自動パイロット**
  - **説明の目的は，操作者が自律システムをどういう状況でどう使うかを判断し，次のタスクを決定するための根拠の提供**
  - **具体的に，** ArduPilot/Software in the Loop (SITL) environment**を想定**

    ⇒実行後にその行動理由を説明するという設定

    ハイレベルとローレベル両方のプラン，判断，制御の説明を含むことが求められている

※ 一般的に，機械学習は教師あり学習，教師なし学習，強化学習の3つに分類され，分類問題は教師あり，教師なしにまたがる．

1.2 説明とは

説明とは，モデルの特徴を意味的な情報と連携させること．
説明戦略として，例えば以下の3つが挙げられる．
⇒哲学的定義における内包的，外延的でいけば，前2者が内包的，最後の1つが外延的といえるか

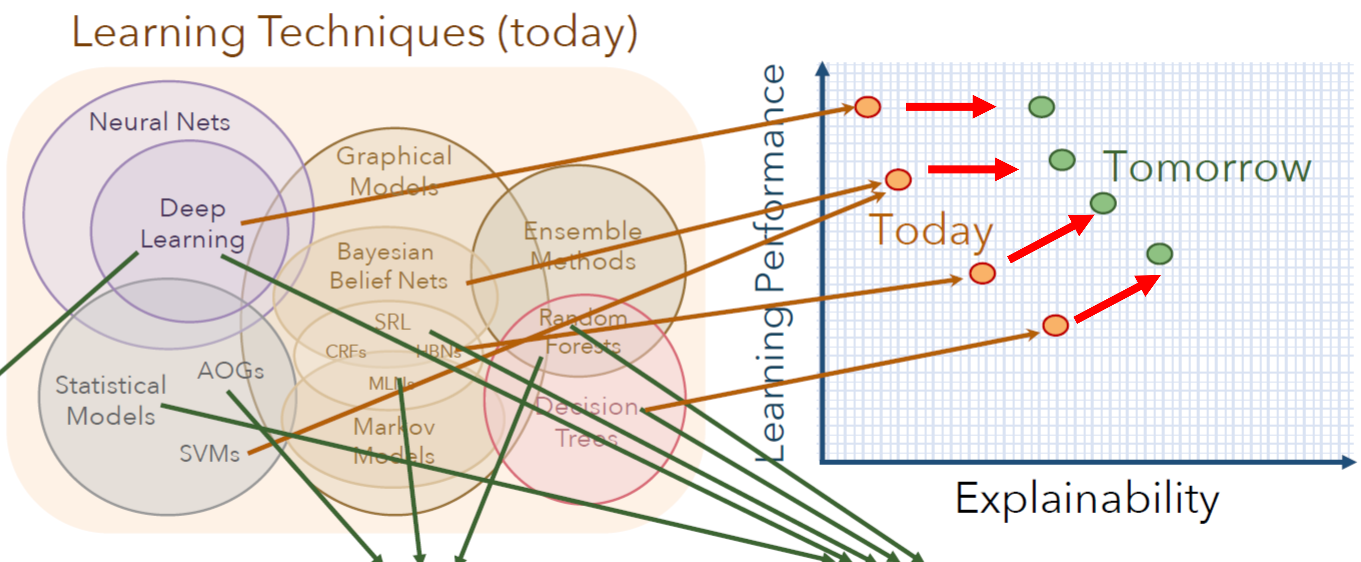
- **Deep Explanation**
  - Deep Learning**向け．**DNN**を**more explainable**にする ⇒完全とは言っていない**
  - DNN**に部分部分を見せて個別に学習させて合成させるなど**（attention mechanisms**や** compositional generative models[IJCAI XAI]）

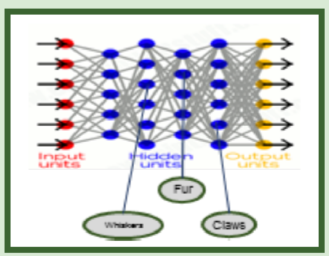
  ⇒どの特徴が判別に効いているのかを示すだけでも分析官の判断を助ける意味では説明足り得る
- **Interpretable Models**
  - Random Forest**や**Bayesian，Probabilistic Logic**など向け**
  - **一般的に**DNN**より表現力，精度は下がるが，ネットワーク内のノードの意味を捉えやすく，モデルの構造や入出力間の相関関係を理解しやすい**
- **Model Induction**
  - **モデル非依存（モデルをブラックボックスとした）手法**
  - **モデルの入出力をより簡単で解析可能なモデルで再現する**（additive feature attribution methods**など**）．**あるいは，別のモデルで説明を生成する**（caption generation）**など**
- **ローカル説明（個々の解釈）とグローバル説明（**AI**の振る舞いの背後にあるロジック）**

1.2 説明とは

XAI Goal

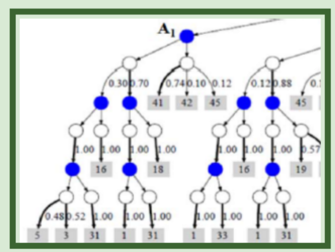
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance





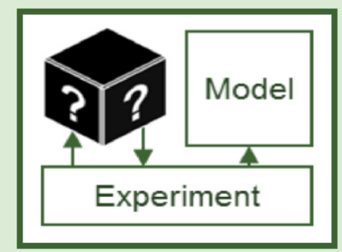
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

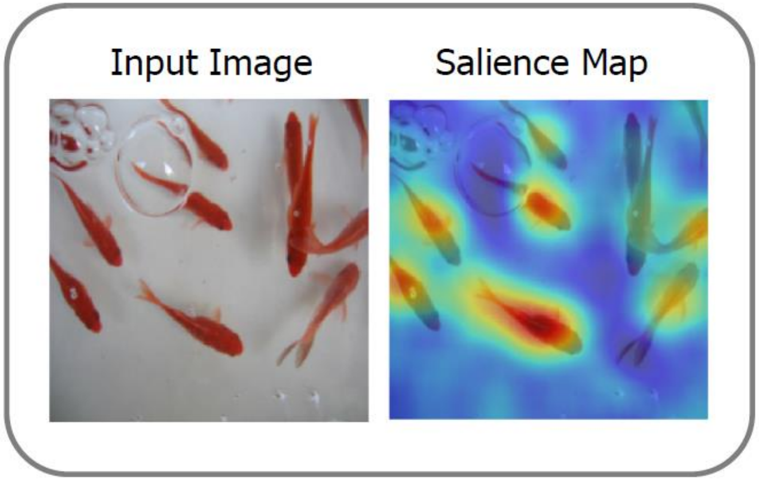


Model Induction

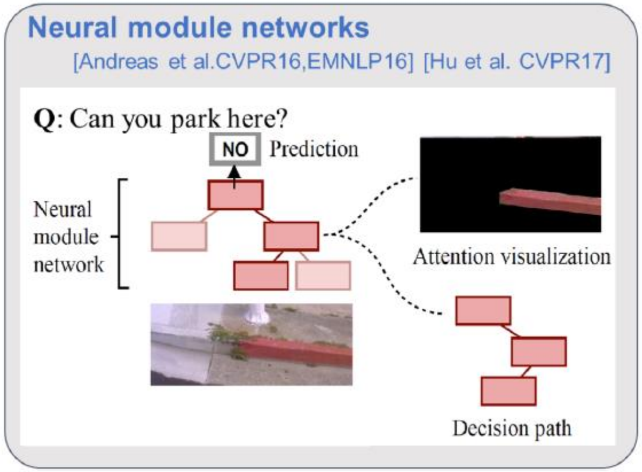
Techniques to infer an explainable model from any model as a black box

1.2 説明とは

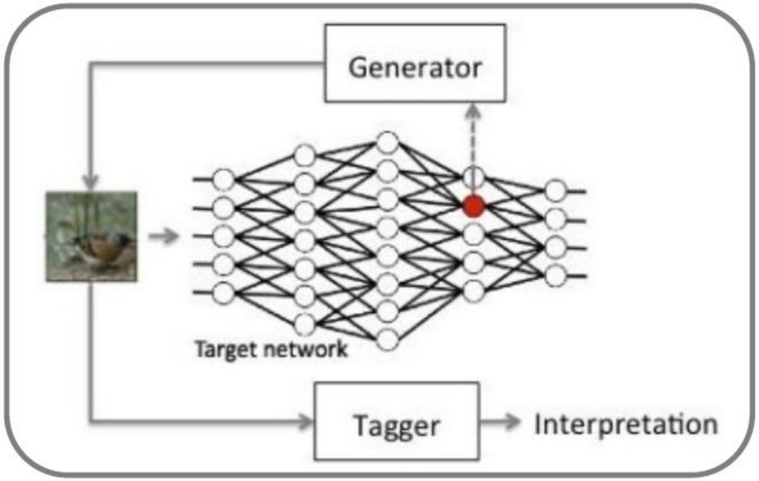
Attention Mechanisms



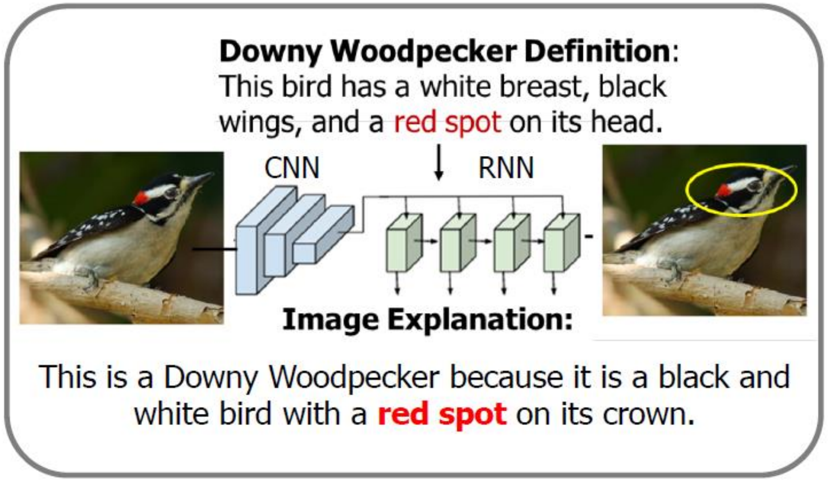
Modular Networks



Feature Identification



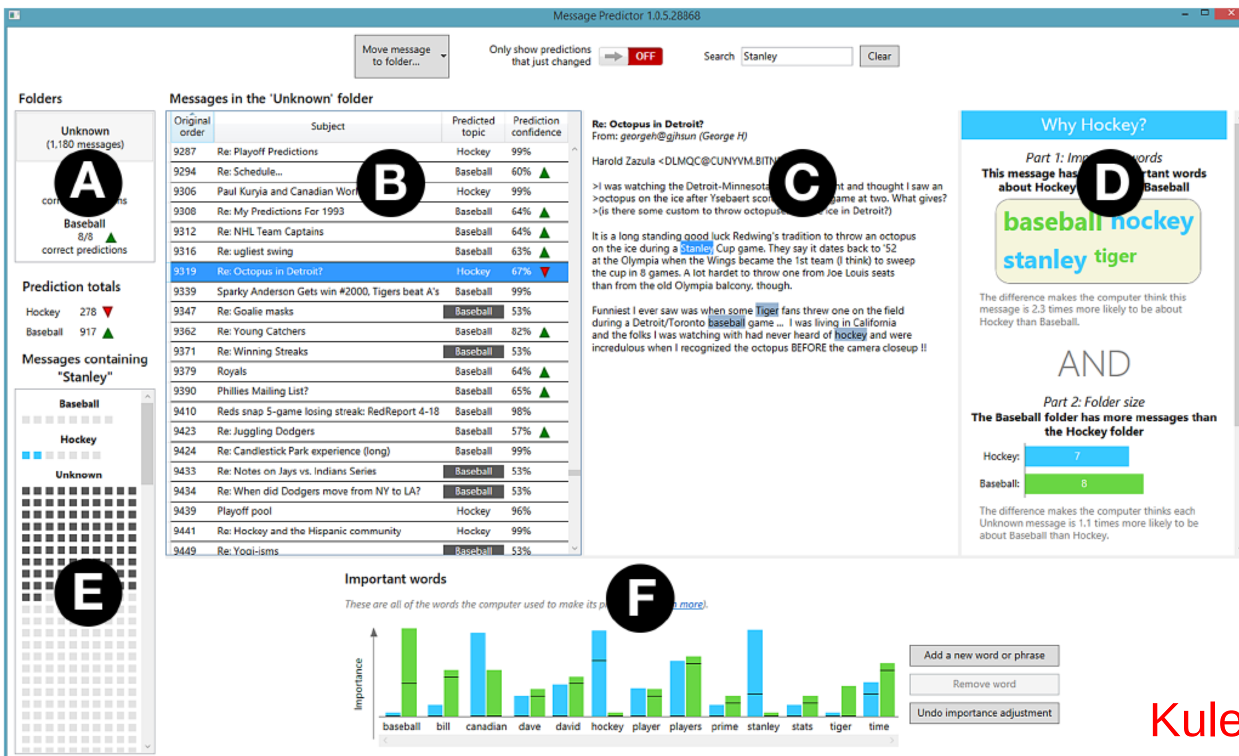
Learn to Explain



1.3 インタフェース

- **詳細は定義されていない**
  - **例やアナロジーによる説明，可視化，言語理解，ダイアログの活用など**
- **最新のHCI技術と認知科学の融合によって理解可能な説明を提示する**

⇒初め から説明戦略とセットで考えるように設定されており，
両者を統合することでブレイクスルーがあると強く主張されている



Kulesza, 2015より

Figure 1. The EluciDebug prototype. (A) List of folders. (B) List of messages in the selected folder. (C) The selected message. (D) Explanation of the selected message's predicted folder. (E) Overview of which messages contain the selected word. (F) Complete list of words the learning system uses to make predictions.

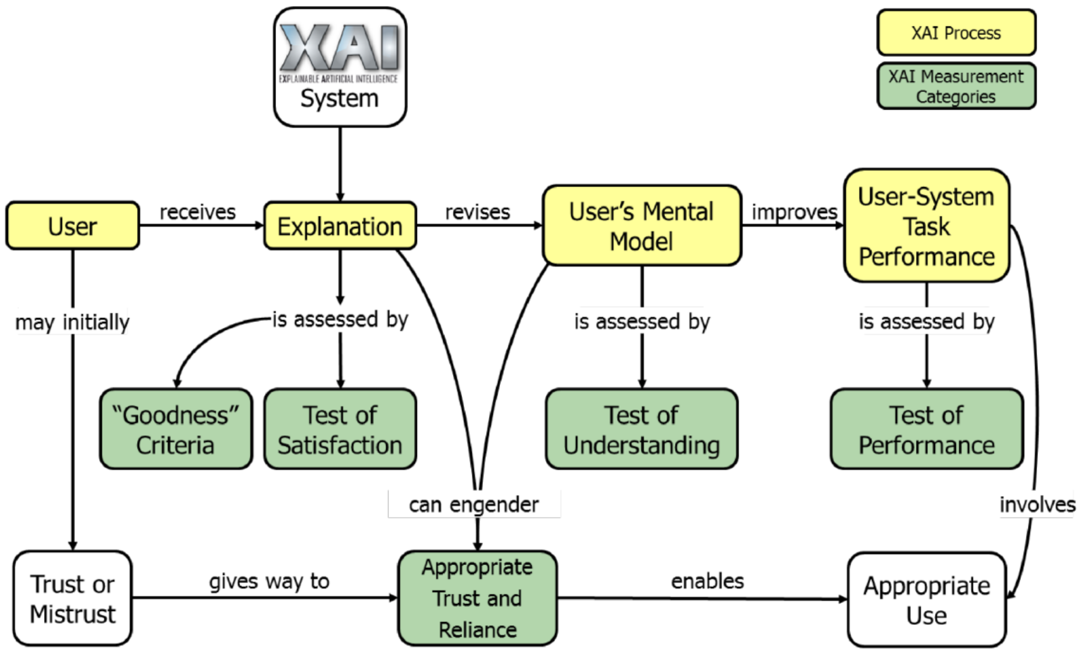
1.4 説明の心理学

- 詳細は定義されていない
- 最新の“説明”の心理学的な定理を拡張し，計算可能な定理を開発する

  ⇒philosophyではなく，psychologyである点が興味深い

  説明の効果を予測するために，計算可能なモデルに特に興味がある．

Explanation Process & Measures



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

1.5 関係しないテーマ

- 一方で，直接的に関係しないテーマには興味がないと明言
  - ユーザモデリング，パーソナライズ，心の定理，インタラクティブ機械学習，可視化解析など

1.6 プロジェクト評価方法

- 学習用データあるいは環境は提供される
  - 但し，実験において人手による大規模な知識構築は避けるべきとされている ⇒戦場での運用が困難なためか？
- 評価は精度と説明の効果の両方
- 応募者は評価者と一緒になって，評価方法と評価指標を決める
- 説明の心理学のチームもアドバイスする

⇒ユーザ満足度はユーザーレイティングによる

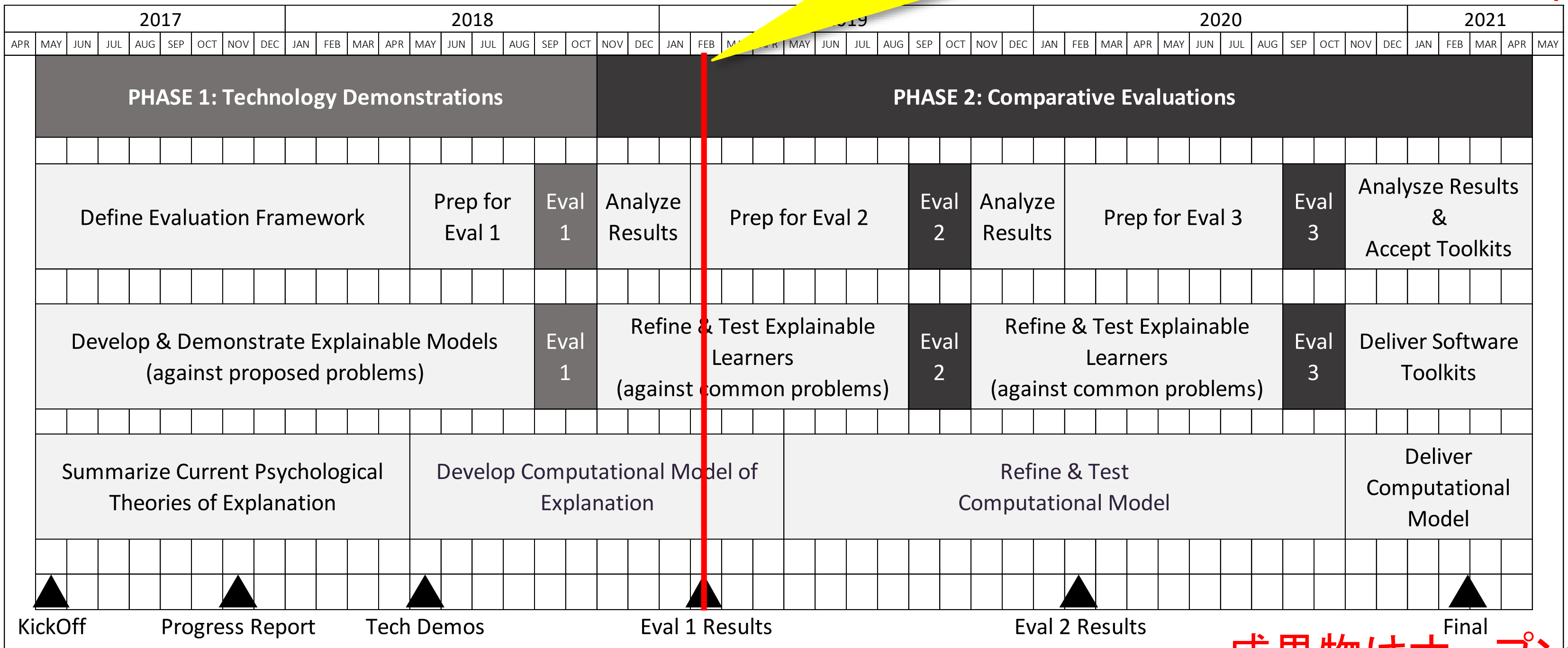
説明の効果測定

User Satisfaction
• Clarity of the explanation (user rating) • Utility of the explanation (user rating)
Mental Model
• Understanding individual decisions • Understanding the overall model • Strength/weakness assessment • 'What will it do' prediction • 'How do I intervene' prediction
Task Performance
• Does the explanation improve the user's decision, task performance?
Trust Assessment
• Appropriate future use and trust
Correctability (Extra Credit)
• Identifying errors • Correcting errors

1.7 スケジュール

現時点で2月のプロジェクト評価
の結果を見つけられず…

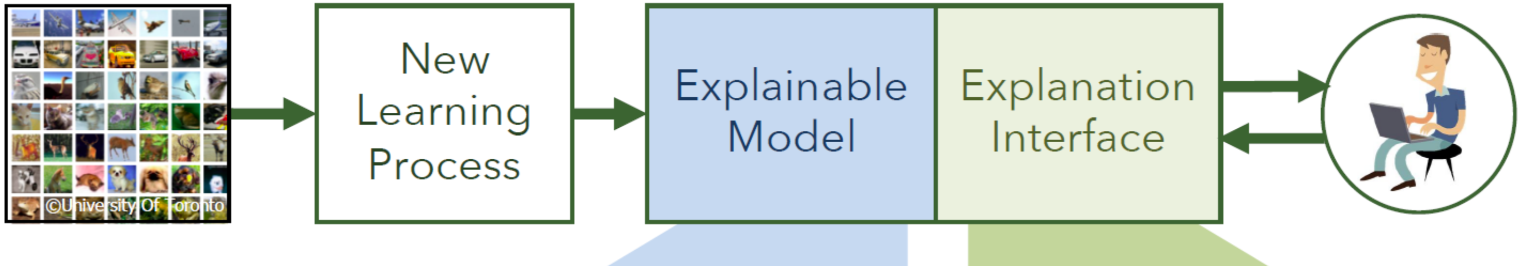
5年間

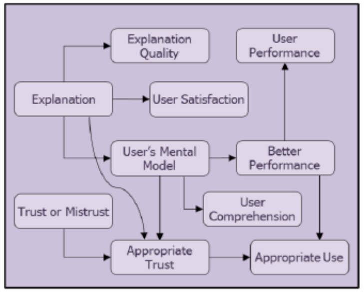


成果物はオープン

- P1では学生などを対象に個別テストを実施，以降よりDoD寄りの共通問題でコンペする
- 2チーム体制
  - TA1: Explainable Learners（説明可能モデルと説明インタフェース）
    - ✓ 11チーム，チーム毎に$800K-$2M/年
  - TA2: Psychological Model of Explanation（説明の心理学）
    - ✓ 1チーム
  - FY2019のXAI予算は$26.05M. トランプ政権下でもAI研究の予算は増加

2. XAIプロジェクト研究





IHMC
Psychological Models of Explanation
Institute for Human & Machine Cognition
Naval Research Laboratory

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

11チーム中深層学習は3チーム

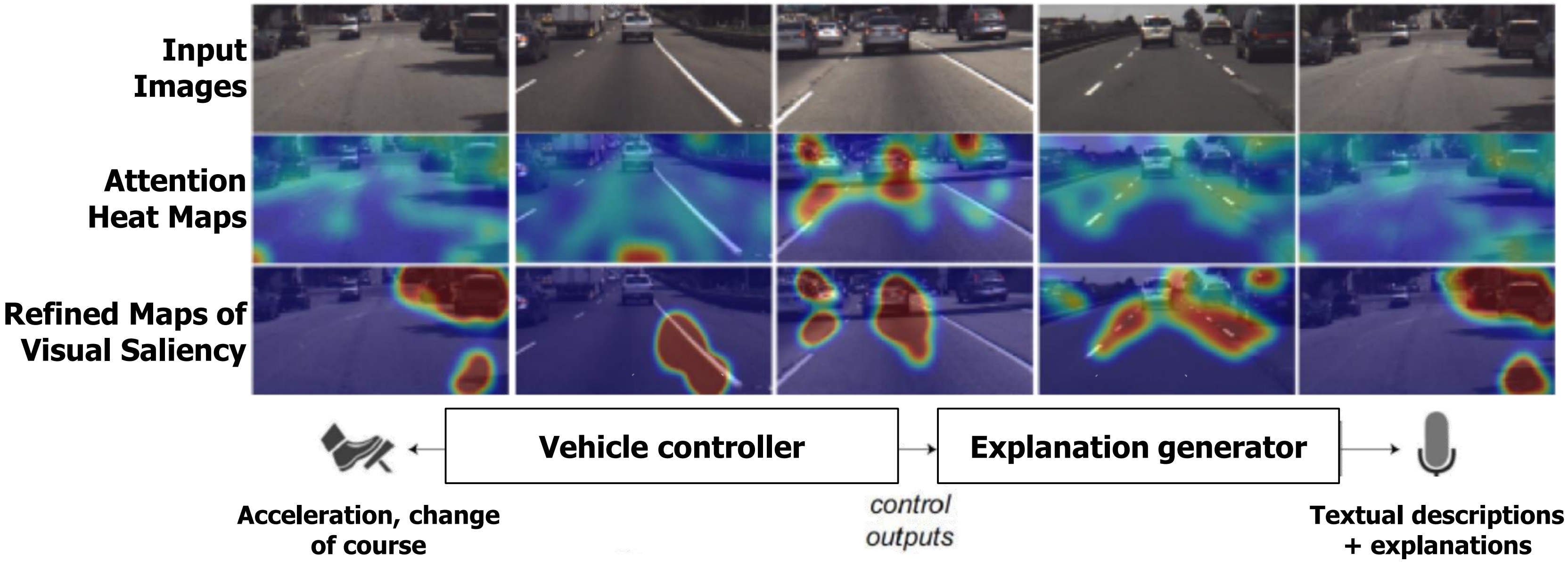
※ Autonomyは状況判断＋計画作成（プラニング）
Analyticsより難しい上にDLはAnalytics向き. UCB要注目

2.1 各プロジェクトの概要

- **UC Berkleyらによる自動運転**
  - **自動運転における判断の適切な説明をヒートマップとテキストで説明する**
- **CRAらによる因果関係モデル**
  - **人工知能が学習したデータと人工知能が出した結論とを合わせて取り込み，その因果関係から人が理解できる理由の説明を行うための因果関係モデルの生成を目指す．**
- **Xerox PARCらによるCOGLE**
  - **人間の概念と機械の学習能力との間に共通領域**(common ground)**を作ることによって，人工知能の決定の意味や将来の振る舞い予測のための情報を人に提供し，自律型ドローンでテストを行う．**
  - **オントロジーを用いてドローンや医療などの分野で使う用語を共通領域を定義．XAIシステムは共通領域を使うことで機械学習モデルの中で起きたことを人の言葉で説明する**
- **Texas A&MらによるFake News Detector**
  - SNS**やニュースなど大量のテキストデータから虚偽情報やフェイクニュースを特定するようモデルを学習させ，どの点でフェイクだと判断したかを述べさせる**
- CMU**と**Stanford**による**XRL
  - XRL(Explainable RL)**では，深層学習について視覚的な説明を助ける高品質なサリエンシー（顕著性）マップを生成する．サリエンシーマップによって**AI**が決定で使った情報や重要性を、視覚的に把握することができる**
- SRI**インターナショナルらによる**DARE
  - **複数の深層学習技術に対応し，**AI**の思考過程を視覚化，決定の説明となる証拠を提供したり，自然言語での説明を生成する**

2.1 **UC Berkley - Deeply Explainable AI for Self-driving Vehicles**

Textual justification system embedded into refined visual attention models to provide appropriate explanation of the behavior of a deep neural vehicle controller



Examples of Action Description and Justification

Action Description	Action Justification
The car accelerates	**because** the light has turned green
The car accelerates slowly	**because** the light has turned green and traffic is flowing
The car is driving forward	**as** traffic flows freely
The car merges into the left lane	**to** get around a slower car in front of it

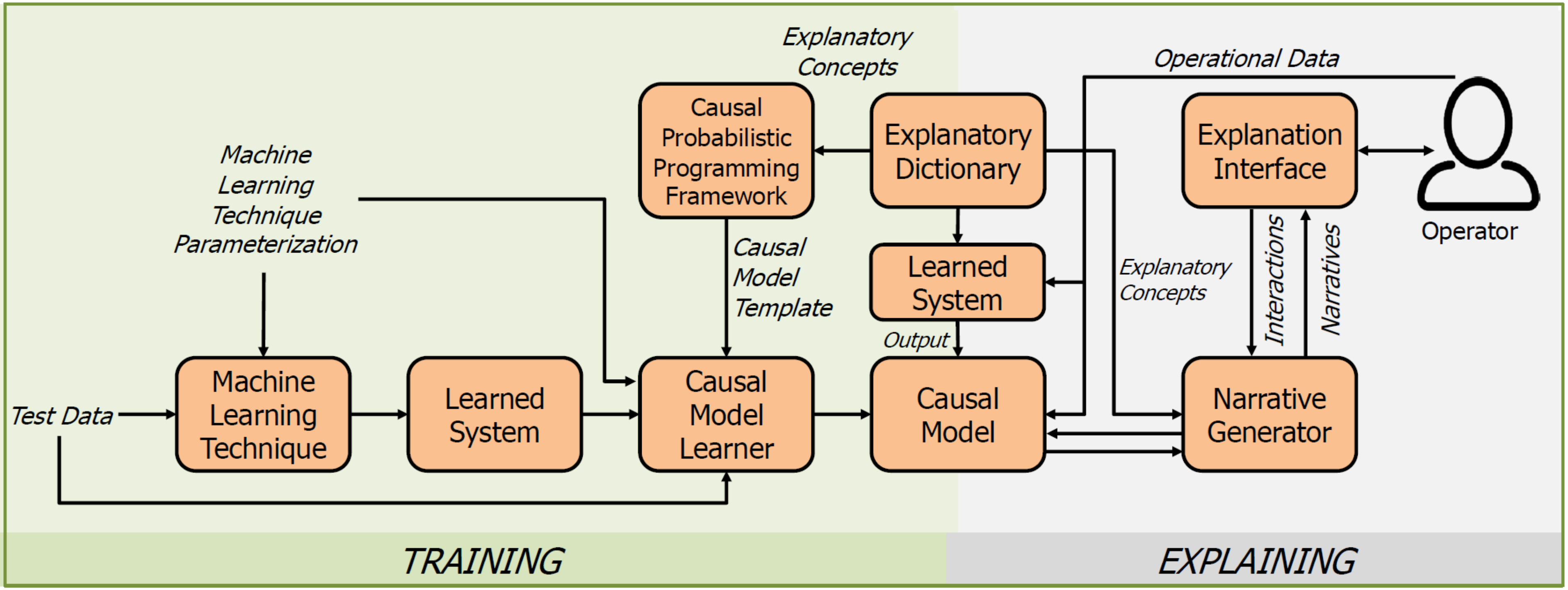
Without explanation: *"The car heads down the street"*
With explanation: *"The car heads down the street because there are no other cars in its lane and there are no red lights or stop signs"*

- Refined heat maps produce more succinct visual explanations and more accurately expose the network's behavior
- Textual action description and justification provides an easy-to-interpret system for self-driving cars

Kim, Rohrbach, Darrell, Canny, and Akata. Show, Attend, Control, and Justify: Interpretable Learning for Self-Driving Cars
Kim and Canny. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention

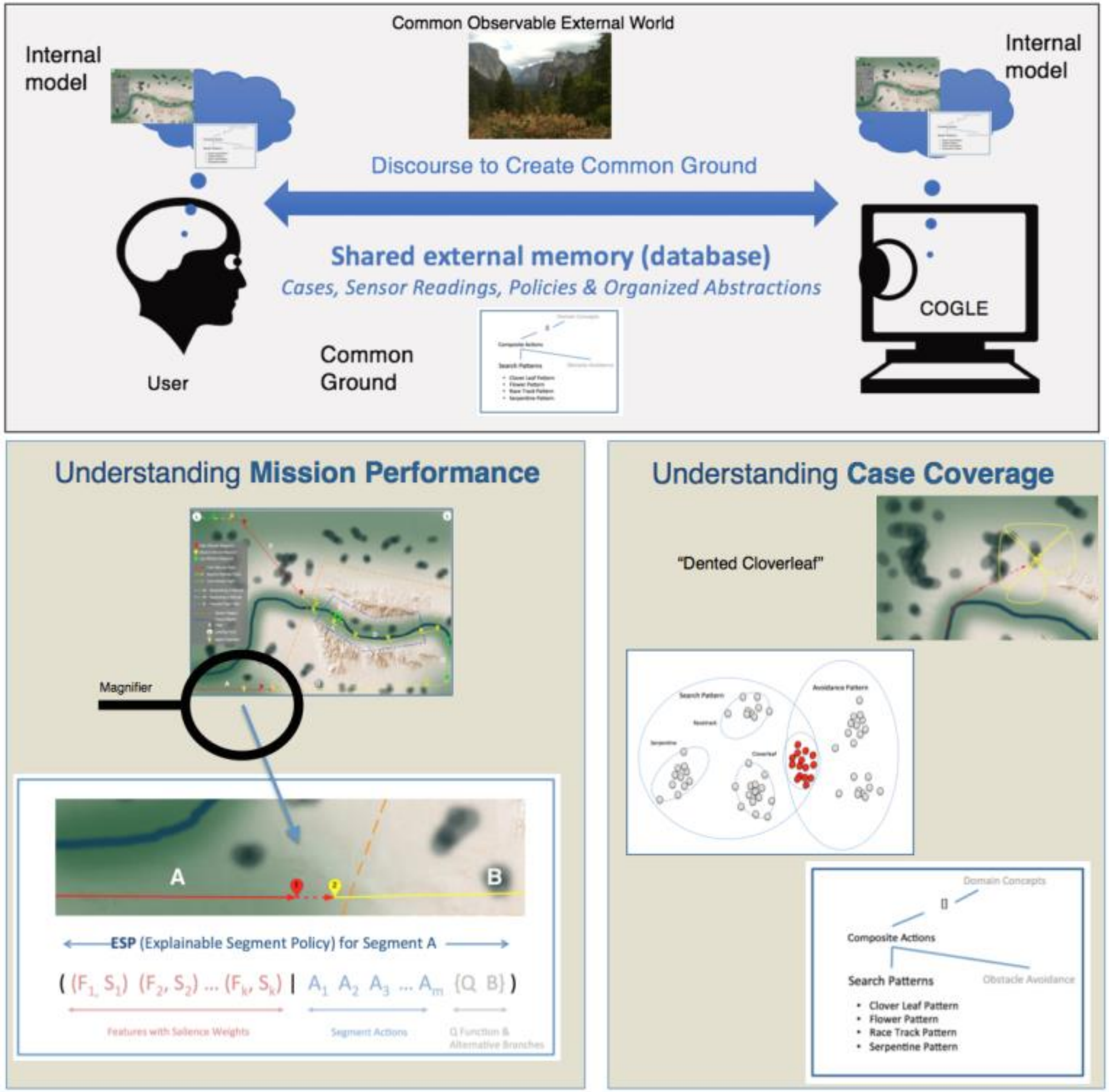
2.2 **CRA -** **Causal Models to Explain Learning (CAMEL)**

Generate causal explanations of ML operation and present them to the user as intuitive narratives in an interactive, easy-to-use interface grounded in cognitive engineering theories



In the model induction approaches, CRA treats the machine learning system as a black box and, their explanation system will run millions of simulation examples and try all sorts of inputs, see what the output is of the system and see if they can infer a model that can describe its behavior. And then they express that model as a probabilistic program, which is a more interpretable model, and use that to generate explanations.

2.3 PARC - COGLE: Common Ground Learning and Explanation

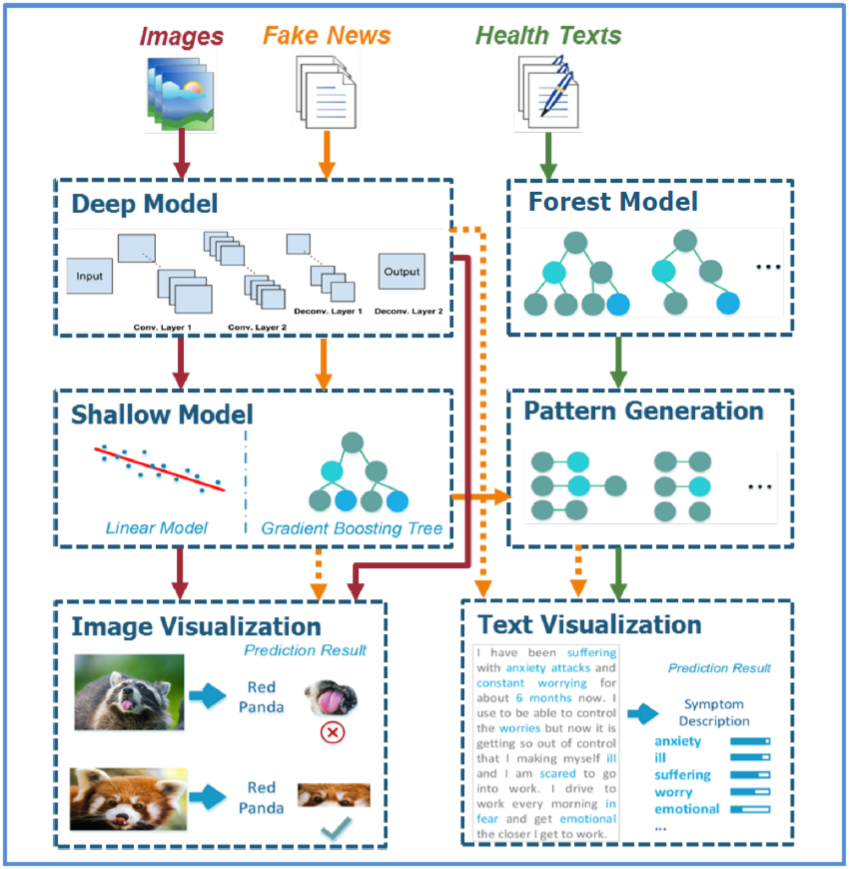


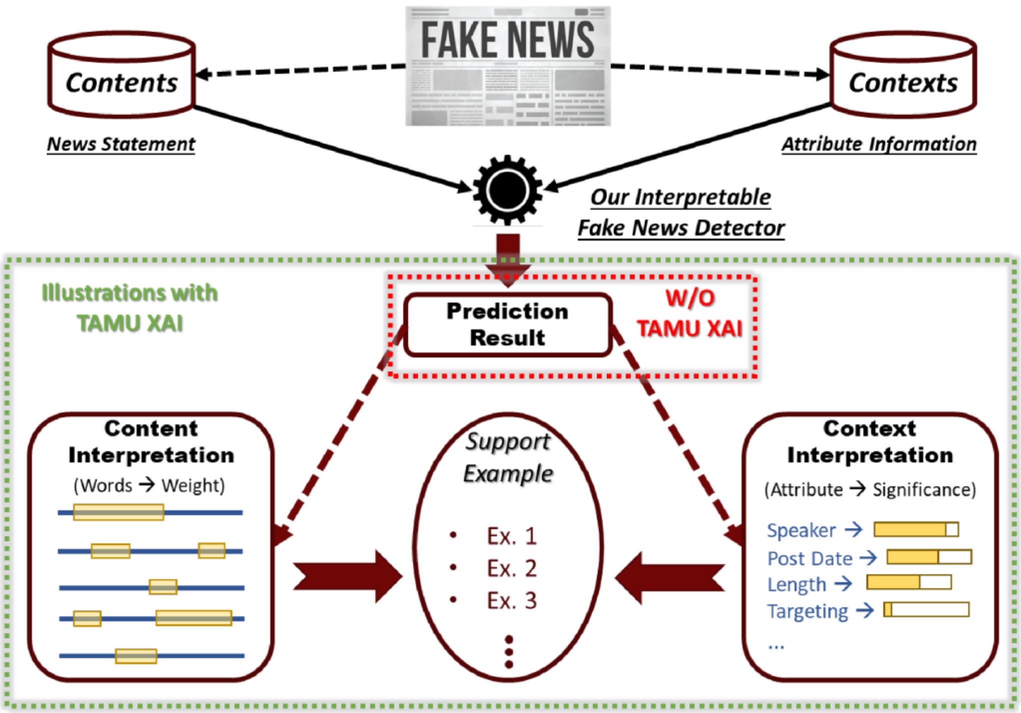
COGLE is developed using UAV test bed that uses reinforcement learning (RL), enabling common ground between people and machine-learning systems, rather than requiring computers to master natural language.

A shared database as external memory for common ground includes actions, domain features, goals and also abstractions of these.
By supporting the creation of common ground, COGLE's explanation interface provides users with explanations and insights into COGLE's reasoning.

2.4 Texas A&M - Transforming DL to Harness the Interpretability of Shallow Models

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





3. 補足 (1/2)

- **品質保証という言い方はしていない．**
  - XAIでは，AIの判断理由（つまり説明）を知ることができれば，正しく動いているかどうかは人間が判断できるという考え方[AAAS]

    ⇒品質保証に繋がる
  - 但し，説明は心理学の側面からアプローチされているため，結局はどれだけ分かった気にさせるかということか？（現実的だが，品質保証の観点では？）
  - そもそも，人間の脳でも“分かった”という状態は分かっていない…

JSAI2018学生セッション（溝口先生 vs. 松尾先生）

分かる（理解する）とは何か？

RNNの出力をある種のシンボルに蒸留できればわかったことになる（松尾）

つまり，ニューロンの一状態と割り切るのか？

具体物がある場合はまだいいが，愛，人情とかはどう学習する？

哲学的に“分かる”ことを理解することは目指さないのか？

分かるということは，永遠に分からないのか？ などなど

3. 補足 (2/2)

- 説明可能性に並ぶ機械学習のもう1つのトピック，公平性は範疇に入っていない
  - 一般的に機械学習アルゴリズムは平均的な損失を最小化するため，マイノリティを無視しても，マジョリティ(の精度)を重視する(当然)
  - 本質的に現代社会における公平性の考え方(アファーマティブアクションのような)にはそぐわない
  - この点に関しても，AIが理由を明らかにすれば人間が公平性を判断できるという考え方か？

- 軍事利用が前提であるため，完全自動化(＝自律殺戮システム)は謳っていない（DoDとしてはあくまで人間の監督下において使うとしている）
  - 初めのwarfighters云々の出だしからも分かるように，まずは意思決定の判断根拠(Informativeness)として利用できるかにフォーカスしている
  - DARPAとしてはAssured Autonomy PJも実施中(AIシステムの振る舞いを特定の範囲に収める)
  - 但し，成果の商用化は謳われている(その中では自動化もあり？線引きは？)
  - AIの軍事利用に関する議論は絶えない(GoogleやAmazonでは反発も)

A. 参考文献

- Explainable Artificial Intelligence (XAI), https://www.darpa.mil/program/explainable-artificial-intelligence, https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf
- Inside DARPA’s effort to create explainable artificial intelligence, https://bdtechtalks.com/2019/01/10/darpa-xai-explainable-artificial-intelligence/
- COGLE, https://www.parc.com/blog/explainable-ai-an-overview-of-parcs-cogle-project-with-darpa/
- DARPA XAI Literature Review, https://arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf
- DARPA's research and development budget for fiscal year 2020 saw a sharp increase in investment in AI applied research,
- https://ru.tenco-tech.com/news/3/243.html
- [AAAS] "How AI detectives are cracking open the black box of deep learning". Science | AAAS. 5 July 2017..
- [IJCAI XAI] Explanation and Justification in Machine Learning: A Survey, Or Biran and Courtenay Cotton, IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)
- Brendel, W., & Todorovic, S. (2011, November). Learning spatiotemporal graphs of human activities. In International Conference on Computer Vision (pp. 778-785).
- Gan, C., Wang, N., Yang, Y., Yeung, D. Y., & Hauptmann, A. G. (2015). Devnet: A deep event network for multimedia event detection and evidence recounting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2568-2577).
- Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating Visual Explanations. arXiv preprint arXiv:1603.08507.
- Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015, March). Principles of explanatory debugging to personalize interactive machine learning. In Proc. of the 20th International Conference on Intelligent User Interfaces (pp. 126-137). ACM.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
- Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9(3), 1350-1371.
- Lombrozo, T. (2006). The structure and function of explanations. Trends in cognitive sciences, 10(10), 464-470.
- Lombrozo, T. (2012). Explanation and abductive inference. Oxford handbook of thinking and reasoning, 260-276.
- Maier, M. E., Taylor, B. J., Oktay, H., & Jensen, D. (2010, July). Learning Causal Models of Relational Domains. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta, GA: AAAI Press.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. arXiv preprint arXiv:1602.04938.
- Yu, Q., Liu, J., Cheng, H., Divakaran, A., & Sawhney, H. (2012, October). Multimedia event recounting with concept based representation. In Proceedings of the 20th ACM international conference on Multimedia (pp. 1073-1076). ACM.
- Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European Conference on Computer Vision (pp. 818-833).