УНИВЕРСИТЕТ ЛОБАЧЕВСКОГО

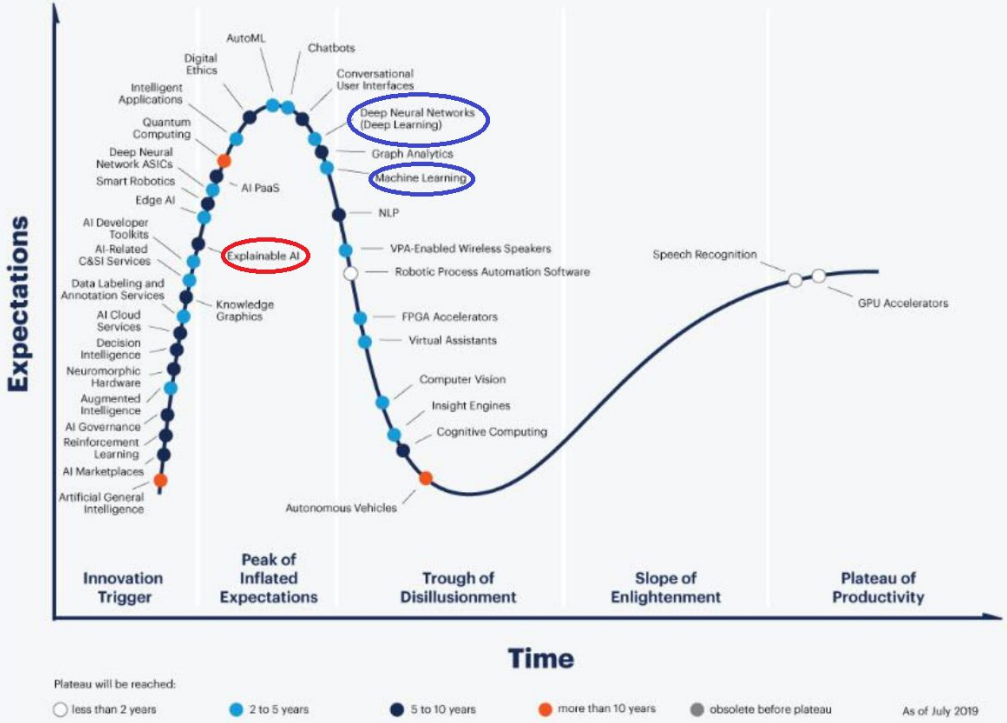
iTmm

Объясняемый искусственный интеллект (XAI)

Н.Ю. Золотых
ННГУ, ИТММ

2 октября 2020

Gartner Hype Cycle for Artificial Intelligence, 2019



gartner.com/SmarterWithGartner

Gartner

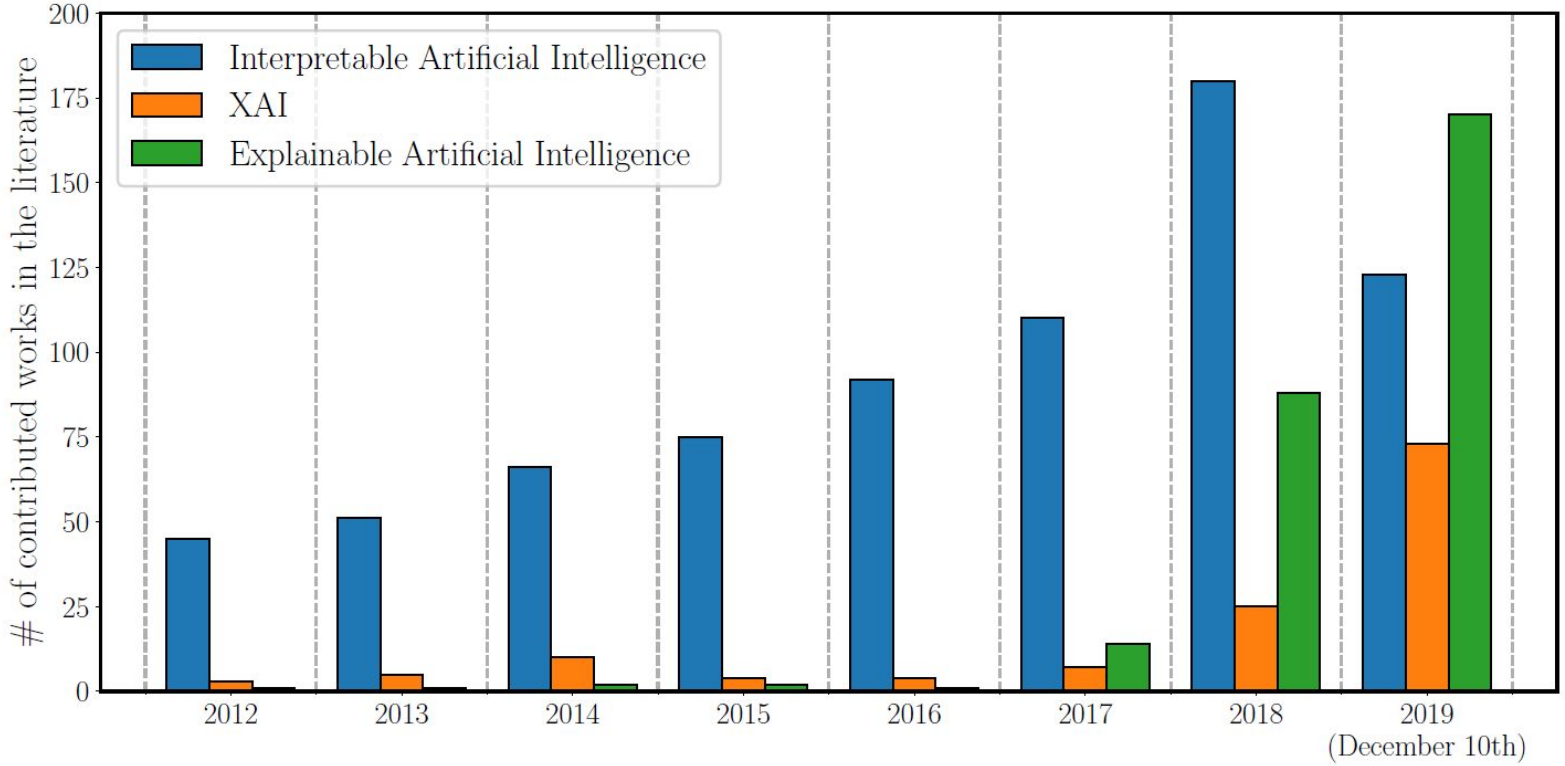


Figure 1: Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus® (December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI models over time (which conforms to intuition, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community.

Source: A. B. Arrieta, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020): 82-115.

- *D. Gunning*, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017)
- *A. B. Arrieta*, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 (2020): 82-115.
- *E. Tjoa, C. Guan*, A survey on explainable artificial intelligence (XAI): Towards medical XAI (2019). arXiv:1907.07374.
- *L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal*, Explaining Explanations: An Overview of Interpretability of Machine Learning (2018). arXiv:1806.00069
- *F. K. Došilović, M. Brćić, N. Hlupić*, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210–215.
- *A. Adadi, M. Berrada*, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
- *O. Biran, C. Cotton*, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), Vol. 8, 2017, p. 1.
- *S. T. Shane, T. Mueller, R. R. Hoffman,W. Clancey, G. Klein*, Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI, Tech. rep., Defense Advanced Research Projects Agency (DARPA) XAI Program (2019).
- *R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi*, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018) 93:1–93:42.

Interpretability & Explainability

Интерпретация - это отображение абстрактного понятия в понятную для человека область [Montavon, Samek, Muller, 2018]

Объяснение - это набор характеристик интерпретируемой области, которые привели к принятию данного решения в конкретном случае [Montavon, Samek, Muller, 2018]

Comprehensibility ≈ interpretability ≈ model-centric explainability

Transparency ≈ explainability ≈ subject-centric explainability

Machine Learning *vs* Data Mining

Говорят, что компьютерная программа *обучается* на основе опыта *E* по отношению к некоторому классу задач *T* и меры качества *P*, если качество решения задач из *T*, измеренное на основе *P*, улучшается с приобретением опыта *E*.

T.M. Mitchell, 1997

Data Mining – совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и *доступных интерпретации знаний*, необходимых для принятия решений в различных сферах человеческой деятельности.

Г. Пятецкий-Шапиро, 1989

Machine Learning *vs* Data Mining

ML и DM извлекают закономерности («знания») из данных, но (немного) с разными целями:

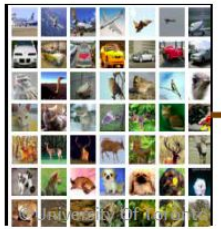
- ML – чтобы обучить машину;
- DM – чтобы обучить человека.

Поэтому *в первую очередь**

- в ML минимизируют ошибку;
- в DM важна интерпретируемость результата.

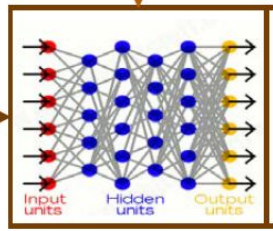
XAI должен стереть эту границу

Today



Training Data

Learning Process



Learned Function

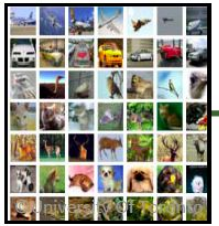
This is a cat
(p = .93)

Output

User with a Task

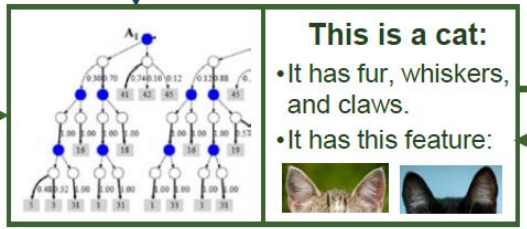
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



Training Data

New Learning Process



Explainable Model

This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

Explanation Interface

User with a Task

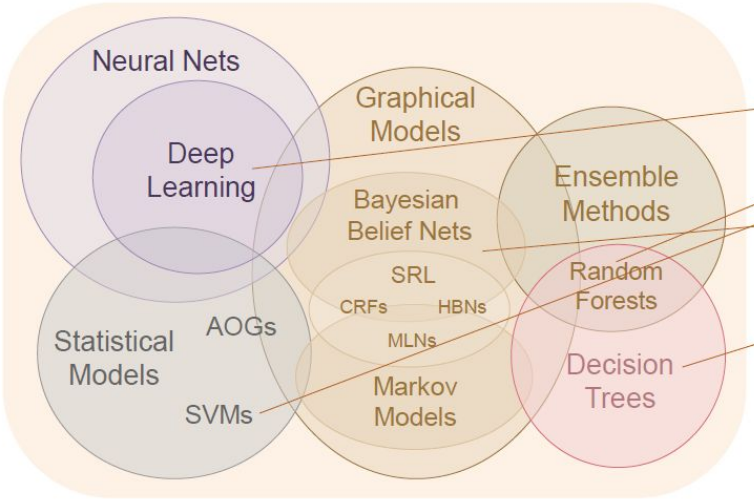
- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Source: D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017)

New Approach

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

Learning Techniques (today)

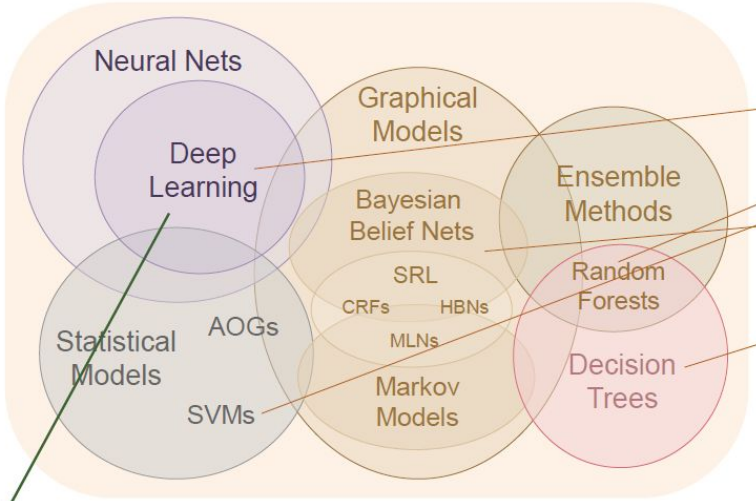


Explainability (notional)

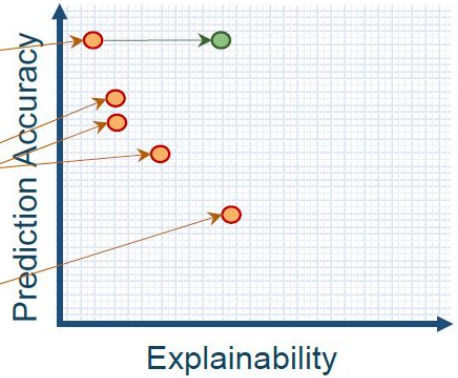
New Approach

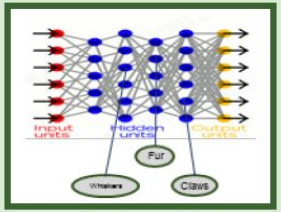
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





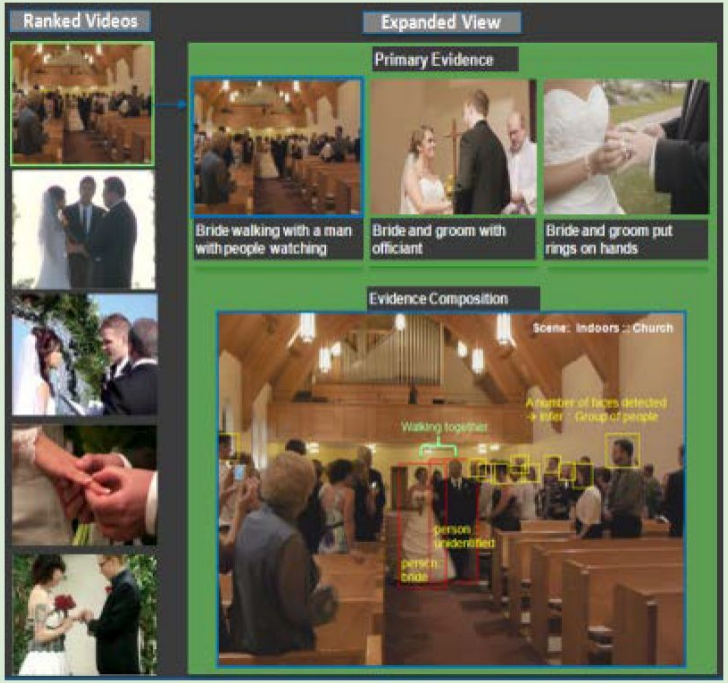
Deep Explanation

Modified deep learning techniques to learn explainable features

Deep Explanation

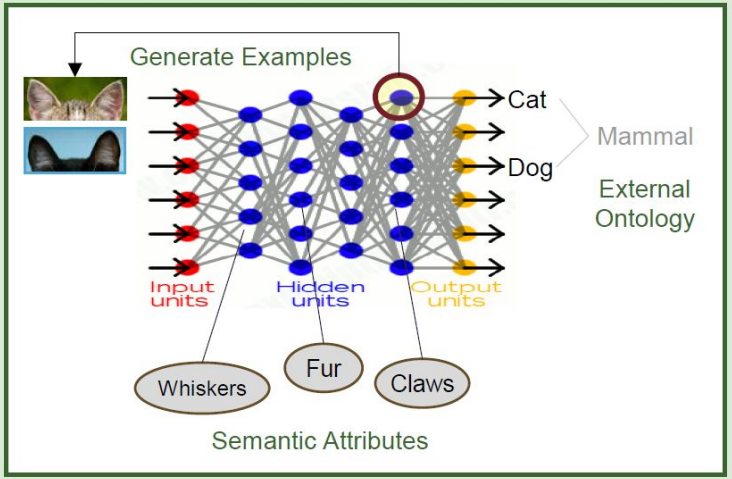
- Извлечение знаний путем комплексного упрощения сети (А.Н. Горбань, 1990, В.Г. Царегородцев, 1998)
- Контрастирование и логически прозрачные нейронные сети (А.Н. Горбань, 1990, Д.И.Еремин, Е.М. Миркес, 1994)
- Визуализация слоев сверточный сети с помощью деконволюции (Zeiler, Fergus, 2014)
- Интерпретация узлов сети как семантических понятий, например, как при идентификации элементов на изображении или регистрации событий (Yu, Liu, Cheng, Divakaran, Sawhney, 2012; Gan, Wang, Ян, Юнг, Хауптманн, 2015)
- Использование методов создания подписей к изображениям (LeCun, Bengio, Hinton, 2015) для генерации объяснений (Hendricks et al., 2016)
- ...

Multimedia Event Recounting



- This illustrates and example of event recounting.
- The system classified this video as a wedding.
- The frames above show its evidence for the wedding classification

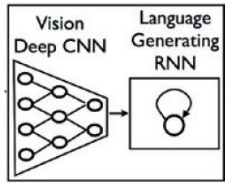
Learning Semantic Associations



- Train the net to associate semantic attributes with hidden layer nodes
- Train the net to associate labelled nodes with known ontologies
- Generate examples of prominent but unlabeled nodes to discover semantic labels
- Generate clusters of examples from prominent nodes
- Identify the best architectures, parameters, and training sequences to learn the most interpretable models

Cheng, H., et al. (2014) SRI-Sarnoff AURORA at TRECVID 2014: Multimedia Event Detection and Recounting.
http://www-nlpir.nist.gov/projects/tvpubs/tv14.papers/sri_aurora.pdf

Generating Image Captions

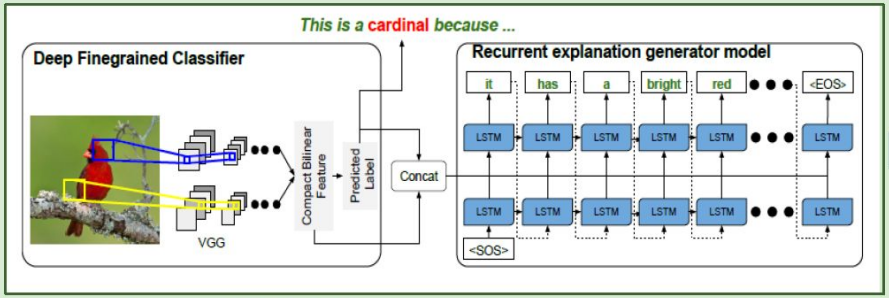


A group of people shopping at an outdoor market

There are many vegetables at the fruit stand

- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions.

Generating Visual Explanations



Researchers at UC Berkeley have recently extended this idea to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

Example Explanations

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

Limitations

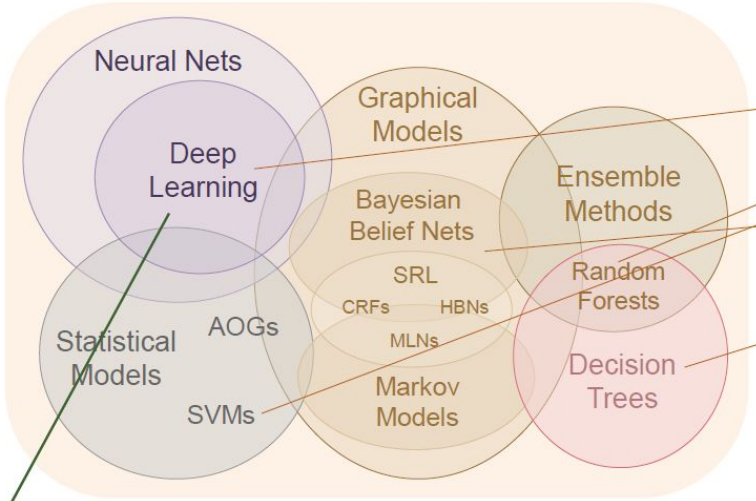
- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors

Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016

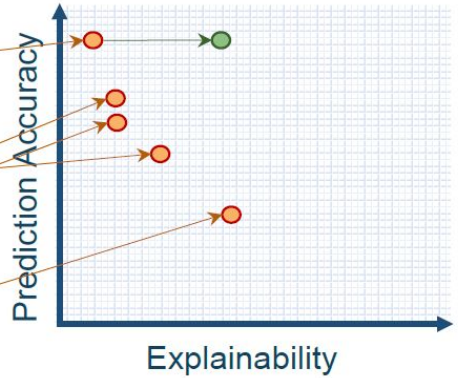
New Approach

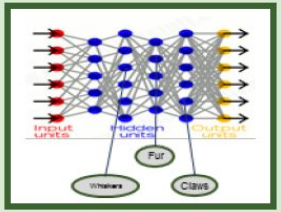
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



Explainability (notional)





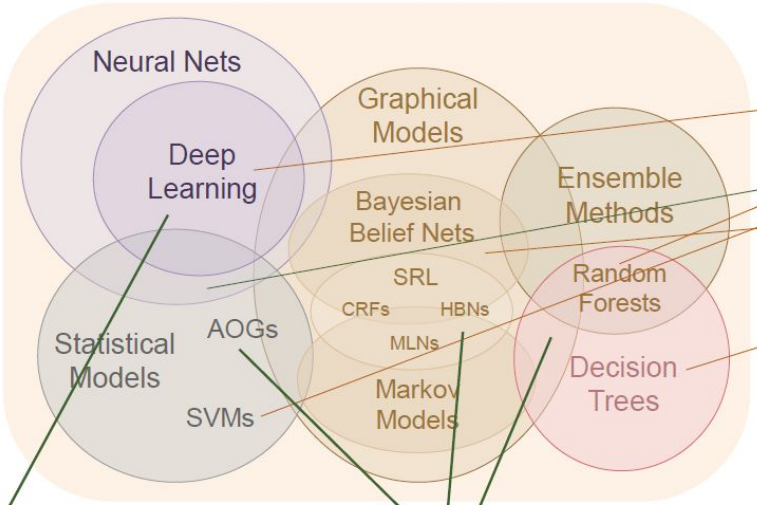
Deep Explanation

Modified deep learning techniques to learn explainable features

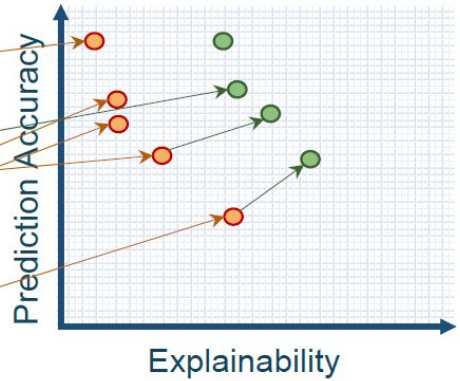
New Approach

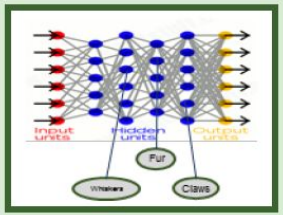
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



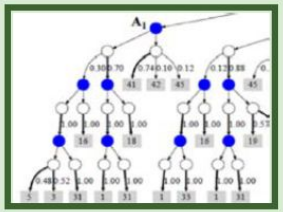
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

Interpretable Models

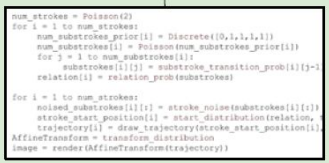
Структурированные, интерпретируемые и причинно-следственные модели:

- Bayesian Rule Lists (Letham, Rudin, McCormick, Madigan, 2015)
- генеративные модели, такие, как Bayesian Program Learning (Lake, Salakhutdinov, Tenenbaum, 2015)
- использование стохастических грамматик (Brendel Todorovic, 2011; Park, Nie, Zhu 2016)
- модели причинно-следственных связей (Maier, Taylor, Oktay, Jensen, 2010)
-

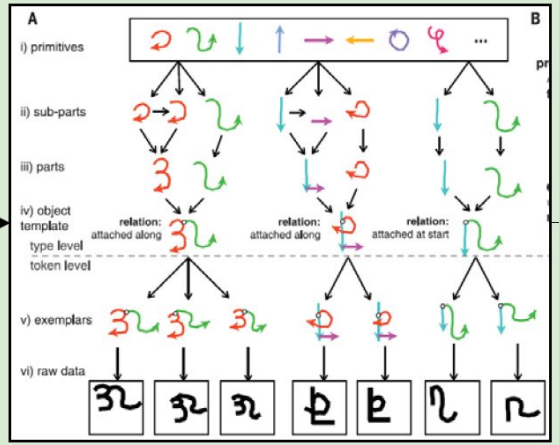
Concept Learning Through Probabilistic Program Induction

Training Data
1623 Characters

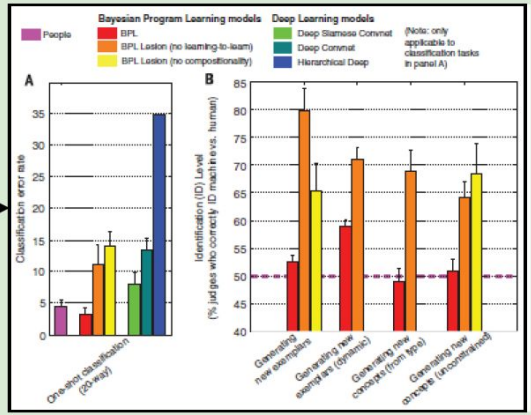
Bayesian Program Learning



Seed Model
A simple Probabilistic Program that describes the parameters of character generation



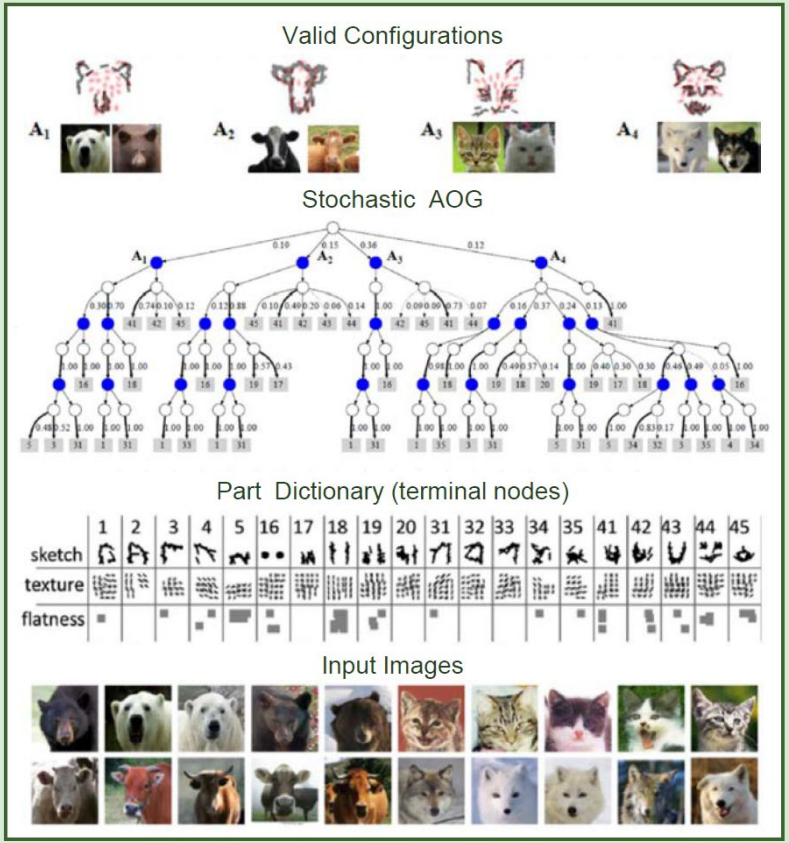
Generative Model
Recognizes characters by generating an explanation of how a new test character might be created (i.e., the most probable sequence of strokes that would create that character)

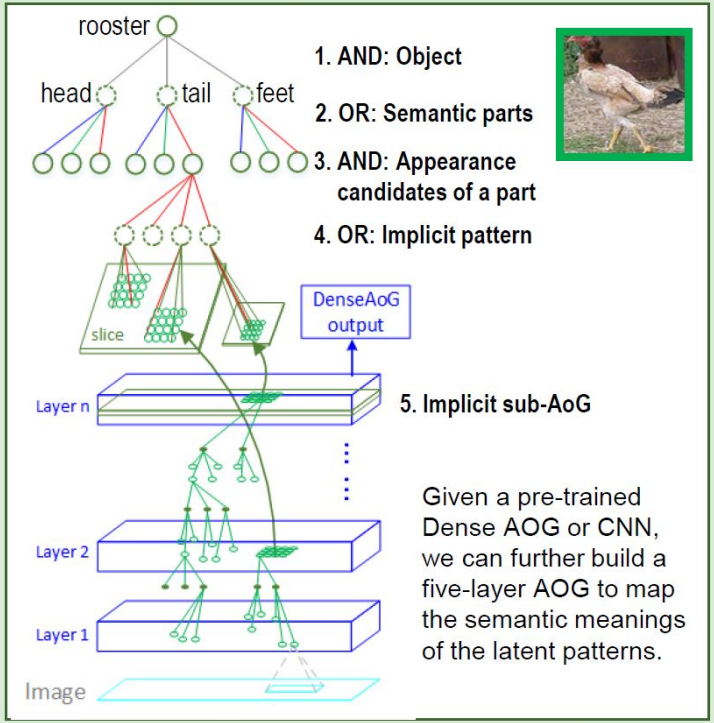


Performance
This model matches human performance and out performs deep learning

Lake, B.H., Salakhutdinov, R., & Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. *Science.* VOL 350, 1332-1338.

Stochastic And-Or-Graphs (AOG)



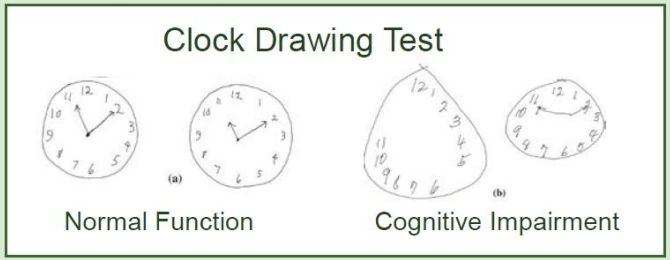


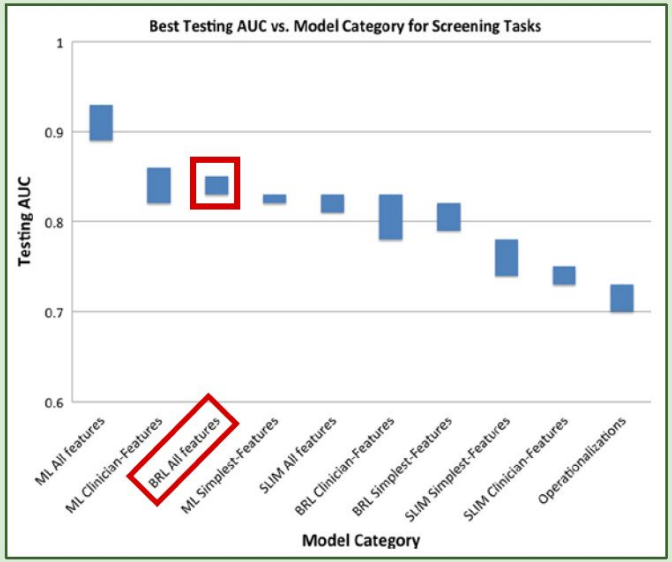
$$L(\theta) = \frac{1}{M}\sum_{m=1}^{M} \underbrace{\log P(I_m, \theta)}_{\text{generative}} + \underbrace{L(pg_m^*, \hat{pg}_m)}_{\text{discriminative}}$$

Si, Z. and Zhu, S. (2013). Learning AND-OR Templates for Object Recognition and Detection. *IEEE Transactions On Pattern Analysis and Machine Intelligence*. Vol. 35 No. 9, 2189-2205.

Bayesian Rule Lists (BRL)

- **if** hemiplegia and age > 60
  - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
  - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
  - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
  - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
  - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
  - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)





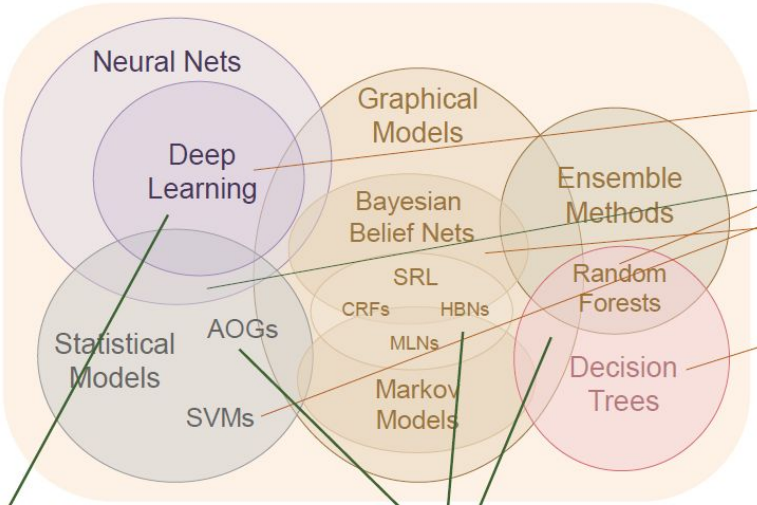
- BRLs are decision lists--a series of if-then statements
- BRLs discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements.
- Experiments show that BRLs have predictive accuracy on par with the current top ML algorithms (approx. 85-90% as effective) but with models that are much more interpretable

Letham, B., Rudin. C., McCormick, T., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics 2015*, Vol. 9, No. 3, 1350-137

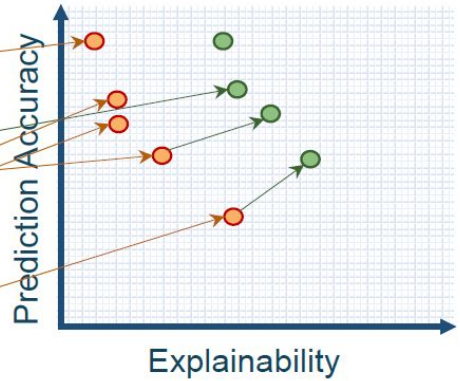
New Approach

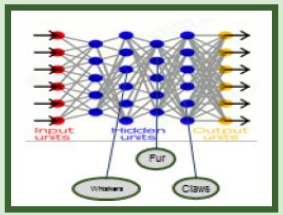
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



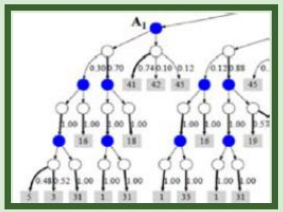
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



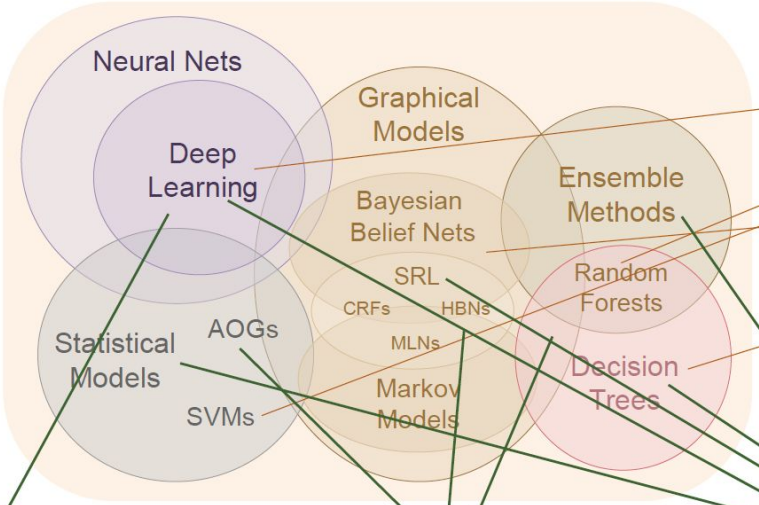
Interpretable Models

Techniques to learn more structured, interpretable, causal models

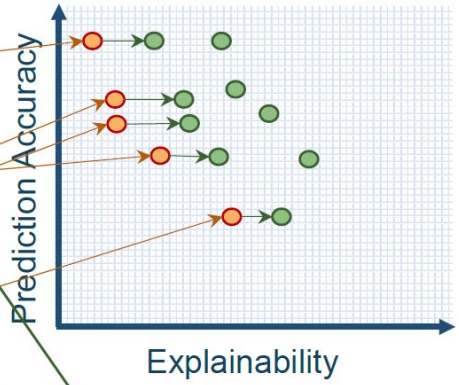
New Approach

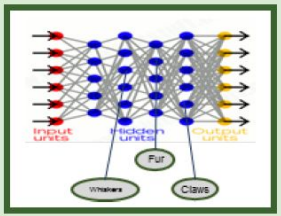
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



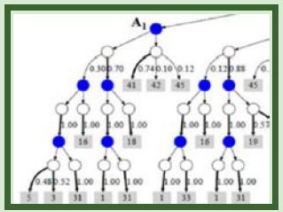
Explainability (notional)





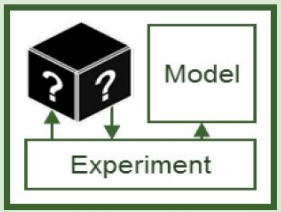
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



Model Induction

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

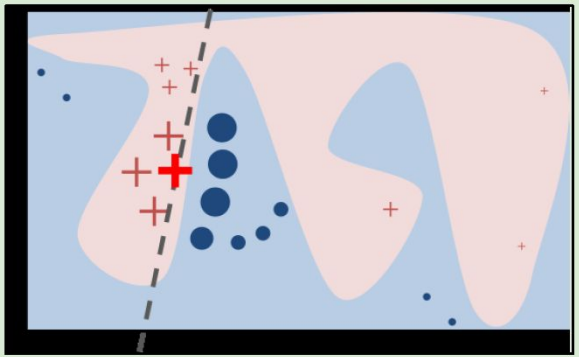
Model Induction

Эксперименты с моделями как с черными ящиками (model-agnostic explanation system).

- исследование откликов на разные входы (например, Ribeiro, Singh, & Guestrin 2016)
- использование абдукции, рассуждений и story generation для “рационализации” правдоподобных объяснений поведения системы
- ...

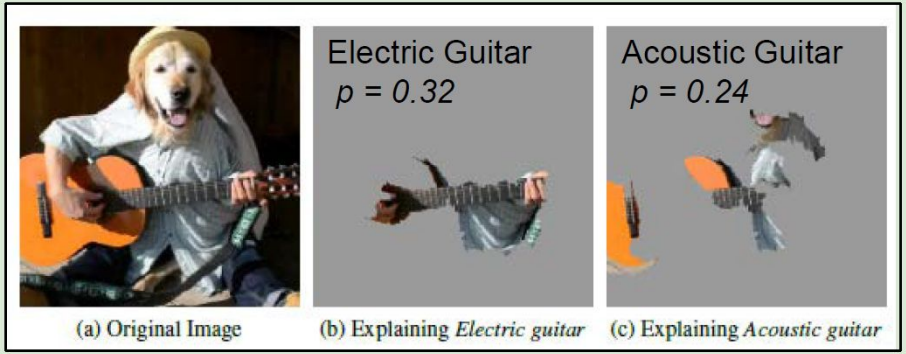
Local Interpretable Model-agnostic Explanations (LIME)

Black-box Induction



The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful. .

Example Explanation



- **LIME** is an algorithm that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model.

- **SP-LIME is** a method that selects a set of representative instances with explanations as a way to characterize the entire model.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *CHI 2016 Workshop on Human Centered Machine Learning. (arXiv:1602.04938v1 [cs.LG] 16 Feb 2016)*

Спасибо за внимание!