CSCE 590-1: Trusted AI

Lecture 24: AI - Unstructured Text – Explanation and AI Testing

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

11TH NOV, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 24

- Introduction Segment
  - Recap from last lecture
  - Paper selection by graduate students

- Main Segment
  - Explanations for text
  - AI testing

- Concluding Segment
  - About next lecture – Lecture 25
  - Ask me anything

Introductory Segment

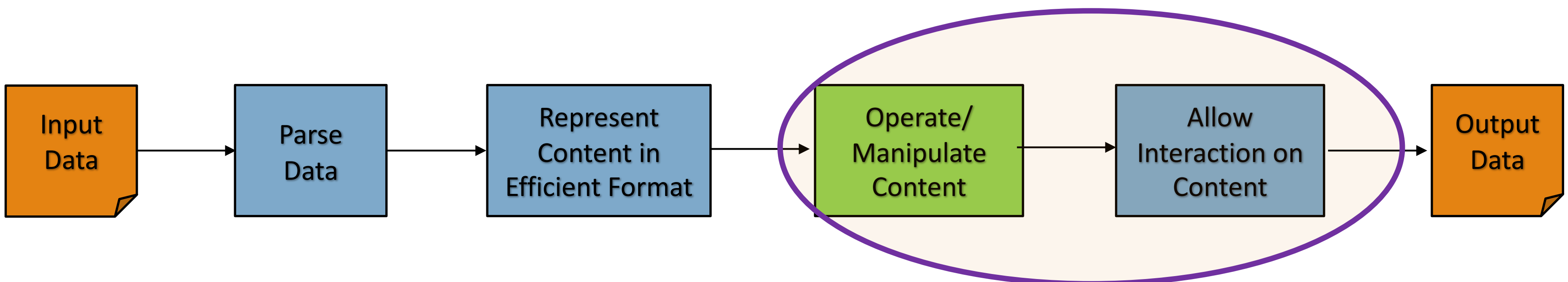
Schedule Snapshot

Oct 26 (Tu)	**Review: Explanation Methods, AIX 360, Discussion**	Quiz 3
Oct 28 (Th)	**Review: project presentations, Discussion**	
Nov 2 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	
Nov 4 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods	
Nov 9 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods	
Nov 11 (Th)	AI – Unstructured Text - Explanation Methods Trust: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	Quiz 4

Recap of Lecture 23

- We looked at rating methods for characterizing machine translators
- We reviewed paper on de-biasing learned word representations

Main Segment



Explanation for Text Classification

LIME – Local Interpretable Model-Agnostic Explanations

Paper: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM's Conference on Knowledge Discovery and Data Mining, KDD2016

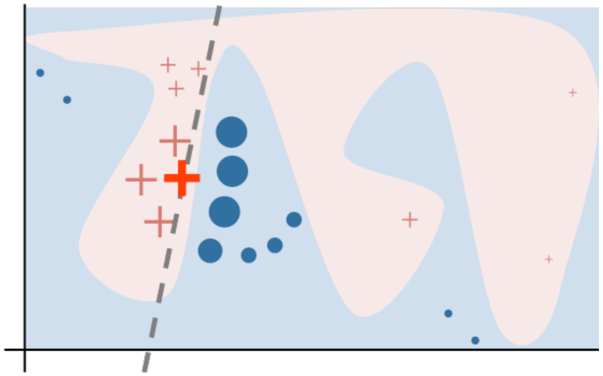
Blogs:

- https://homes.cs.washington.edu/~marcotcr/blog/lime/
- https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Code: https://github.com/marcotcr/lime

Figures credit: Marco Tulio Ribeiro

LIME Key Idea

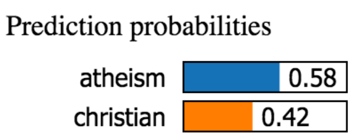


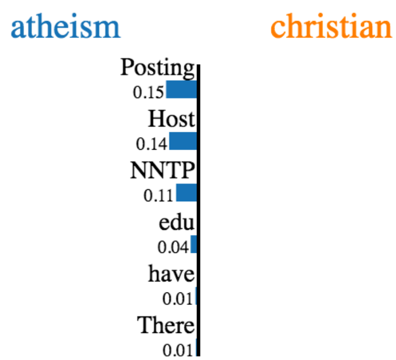
- Generate a local, linear explanation for any model
- How
  - Perturb near the neighborhood of a point of interest, X (**Local**)
  - Fit a linear function to the model's output (**Linear**)
  - Interpret coefficients of the linear function (**Explain**)
  - **Visualize**
- Applicability
  - Any classification model!

LIME on Text

Question: Why is a classifier with >90% accuracy predicting based on ?

Task: classifying religious inclination from email text





Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

"If we remove the words **Host** and **NNTP** from the document, we expect the classifier to predict **atheism** with probability 0.58 - 0.14 - 0.11 = 0.31"

Source: https://github.com/marcotcr/lime

Evaluation of Explanation Methods

- Text
  - Human-grounded Evaluations of Explanation Methods for Text Classification, Piyawat Lertvittayakumjorn, Francesca Toni, https://arxiv.org/abs/1908.11355, 2019

- Image
  - Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey, Arun Das, Paul Rad, https://arxiv.org/abs/2006.11371, 2020

- Many data types (image, text, audio, and sensory domains):
  - **How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods,** *Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani Srivastava,* Advances in Neural Information Processing Systems 33 (NeurIPS 2020), https://proceedings.neurips.cc/paper/2020/hash/2c29d89cc56cdb191c60db2f0bae796b-Abstract.html

One Class, Many Explanations

An example from the Amazon dataset, Actual: Pos, Predicted: Pos, (Predicted scores: Pos 0.514, Neg 0.486):
"OK but not what I wanted: These would be ok but I didn't realize just how big they are. I wanted something I could actually cook with. They are a full 12" long. The handles didn't fit comfortably in my hand and the silicon tips are hard, not rubbery texture like I'd imagined. The tips open to about 6" between them. Hope this helps someone else know ..."

Method	**Top-3 evidence texts**	**Top-3 counter-evidence texts**
LIME (W)	comfortably / wanted / helps	not / else / someone
LRP (W)	are / not / 6	: / tips / open
LRP (N)	are hard , not / about 6" between / not what I wanted	. The tips open / : These would / in my hand and
Grad-CAM-Text (N)	comfortably in my hand / I wanted : These / . The tips open	not what I wanted / not rubbery texture like / Hope this helps someone
DTs (N)	imagined . The tips	'd imagined . / are . I wanted / would be ok

Table 3: Examples of evidence and counter-evidence texts generated by some of the explanation methods.

Image source: Human-grounded Evaluations of Explanation Methods for Text Classification, Piyawat Lertvittayakumjorn, Francesca Toni, https://arxiv.org/abs/1908.11355, 2019

Explanation Evaluation Tasks

	Task 1 (Section 3.1)	**Task 2** (Section 3.2)	**Task 3** (Section 3.3)
Assumption	Good explanations can reveal model behavior	Good explanations justify the predictions	Good explanations help humans investigate uncertain predictions
Model(s)	Two classifiers with different performance on a test dataset	One well-trained classifier	One well-trained classifier
Input text	A test example for which both classifiers predict the same class	A test example which the classifier predicts with high confidence ($\max_c \mathbf{p}_c > \tau_h$)	A test example which the classifier predicts with low confidence ($\max_c \mathbf{p}_c < \tau_l$)
Information displayed	1. The input text 2. The predicted class 3. (Highlighted) top-m evidence texts of each model	1. Top-m evidence texts	1. The predicted class 2. The predicted probability $\mathbf{p}$ 3. Top-m evidence and top-m counter-evidence texts
Human task	Select the more reasonable model and state if they are confident or not	Select the most likely class of the document which contains the evidence texts and state if they are confident or not	Select the most likely class of the input text and state if they are confident or not
Scores to the explanation method	(-)1.0: (In)correct, confident (-)0.5: (In)correct, unconfident 0.0: No preference	(-)1.0: (In)correct, confident (-)0.5: (In)correct, unconfident 0.0: No preference	(-)1.0: (In)correct, confident (-)0.5: (In)correct, unconfident

Table 1: A summary of the proposed human-grounded evaluation tasks.

Image source: Human-grounded Evaluations of Explanation Methods for Text Classification, Piyawat Lertvittayakumjorn, Francesca Toni, https://arxiv.org/abs/1908.11355, 2019

UI to Provide Feedback

	Task 1 (Section 3.1)
Assumption	Good explanations can reveal model behavior
Model(s)	Two classifiers with different performance on a test dataset
Input text	A test example for which both classifiers predict the same class
Information displayed	1. The input text 2. The predicted class 3. (Highlighted) top-m evidence texts of each model
Human task	Select the more reasonable model and state if they are confident or not
Scores to the explanation method	(-)1.0: (In)correct, confident (-)0.5: (In)correct, unconfident 0.0: No preference

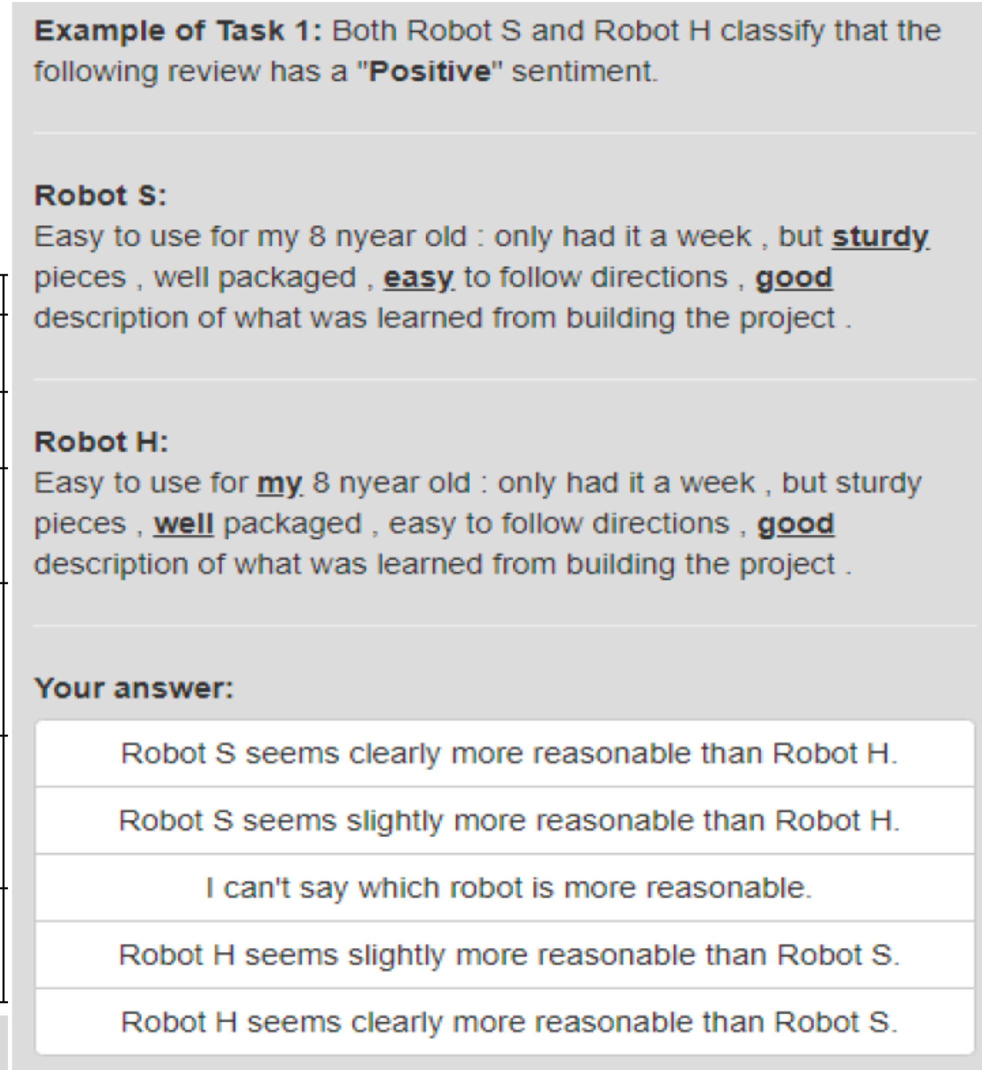


Image sources: Human-grounded Evaluations of Explanation Methods for Text Classification, Piyawat Lertvittayakumjorn, Francesca Toni, https://arxiv.org/abs/1908.11355, 2019

Results of Human Evaluation

Explanation Method	Task 1						Task 2						Task 3					
	Amazon			ArXiv			Amazon			ArXiv			Amazon			ArXiv		
	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘	$\mathcal{A}$	✔	✘
Random (W)	.02	.00	.04	-.11	-.05	-.17	.06	.10	.02	.07	.09	.04	.05	.53	-.43	.01	.32	-.30
Random (N)	.02	.02	.02	-.12	-.16	-.07	.12	.13	.12	.29	.32	.25	.01	.54	-.55	.02	.29	**-.25**
LIME (W)	-.02	.02	-.06	.03	.02	**.03**	**.69**	**.74**	.64	**.70**	**.75**	**.64**	.02	.50	-.45	-.02	.31	-.34
LRP (W)	.00	-.01	.02	-.03	-.01	-.05	.13	.26	-.01	.26	.36	.16	-.02	.50	-.54	-.06	.33	-.44
LRP (N)	-.07	-.04	-.09	**.12**	**.24**	-.01	.26	.45	.08	.44	.49	.39	.08	.60	-.43	**.17**	**.60**	-.26
DeepLIFT (W)	.04	.03	.04	.07	.13	.00	.21	.37	.04	.26	.35	.16	-.03	.47	-.53	-.08	.28	-.44
DeepLIFT (N)	.06	.06	**.05**	.06	.22	-.10	.23	.47	-.01	.38	.47	.28	.05	.59	-.49	.02	.33	-.30
Grad-CAM-T (N)	**.07**	**.11**	.03	-.03	-.04	-.01	.65	.64	**.66**	.53	.65	.41	.05	.51	-.42	.06	.56	-.45
DTs (N)	-.05	-.02	-.08	-.13	-.22	-.03	.64	.68	.59	.51	.69	.32	**.10**	**.60**	**-.40**	-.11	.29	-.50
Fleiss κ (Amazon)	0.050 / 0.054			N/A			0.274 / 0.371			N/A			0.212 / 0.499			N/A		

Table 4: The average scores of the three evaluation tasks. The score range is [-1,1] in which 1 is better. $\mathcal{A}$, ✔, and ✘are for all, correctly classified, and misclassified input texts, respectively. Boldface numbers are the highest average scores in the columns. A number is underlined when there is no statistically significant difference between the scores of the corresponding method and the best method in the same column (at a significance level of 0.05). The last row reports inter-rater agreement measures (Fleiss' kappa) in the format of α / β where α considers answers with human confidence levels (5 categories for task 1-2 and 4 categories for task 3) and β considers answers regardless of the human confidence levels (3 categories for task 1-2 and 2 categories for task 3).

Image source: Human-grounded Evaluations of Explanation Methods for Text Classification, Piyawat Lertvittayakumjorn, Francesca Toni, https://arxiv.org/abs/1908.11355, 2019

AI Testing

Types of Evaluation

- **Technical evaluation**: is the (new) method accurate?
  - Concern is bug-free implementation
- **Business evaluation**: is the (new) method beneficial ?
  - Usually done in a small time horizon
  - Other factors may correlate with success or failure
- **Causal evaluation**: did the (new) method really work?
  - No other factor but this method contributed to business success

AI *for* Testing

- AI for testing
  - Test case and data generation
  - “Value” based testing
- Sample of work
  - **Blogs**: https://www.perfecto.io/blog/ai-in-software-testing ; https://www.testingxperts.com/blog/AI-in-Software-Testing
  - **Papers**: Artificial Intelligence in Software Test Automation: A Systematic Literature Review, Anna Trudova, Michal Dolezel, Alena Buchalcevová, Published in ENASE 2020, https://www.semanticscholar.org/paper/Artificial-Intelligence-in-Software-Test-A-Review-Trudova-Dolezel/ccbe24b348194905edeca784776255007 86e55d6;
    T. M. King, J. Arbon, D. Santiago, D. Adamo, W. Chin and R. Shanmugam, "AI for Testing Today and Tomorrow: Industry Perspectives," *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 2019, pp. 81-88, doi: 10.1109/AITest.2019.000-3.

Mapping AI Methods to Test Activities

Table 4: Mapping AI techniques and testing activities (x = technique is applicable).

AI technique \ Testing activity	Publications identifier	Test case generation	Test oracle generation	Test execution	Test data generation	Test results reporting	Test repair	Test case selection	Flaky test prediction	Test order generation
Non-maximum suppression method (NMS)	R32					x				
SIFT, FAST, and FNCC algorithms	R24, R37						x			
Contour detection, OCR	R37						x			
Bayesian Network	R7, R28	x						x	x	
Particle swarm optimization (PSO)	R18							x		
Hybrid genetic algorithms	R14, R16	x			x					
Ant colony optimization (ACO)	R2, R8	x			x					
Artificial Neural Network (ANN)	R3, R11, R22, R23	x	x							
Graphplan algorithm	R9, R26	x	x							
Support vector machine (SVM)	R36, R38	x	x							
AdaBoostM1 and Incremental Reduced Error Pruning (IREP) algorithms	R33		x							
Convolutional Neural Networks (CNN)	R29		x		x					
Template-matching algorithm	R32, R35		x							
Decision tree algorithm (C4.5)	R4	x								
Markov model	R31	x								
MF-IPP (Multiple Fact Files Interference Progression Planner)	R15	x								
Algorithm from NLP field	R25	x								
Q-learning	R12, R17, R30	x								
Recurrent neural network (RNN)	R13				x					
L*	R39			x	x					
Fuzzing algorithm	R20			x						
k-means	R21			x						
KStar classifier	R19			x						
Heuristics algorithms	R27									x

Testing *for* AI

- Papers
  - A. Aggarwal, S. Shaikh, S. Hans, S. Haldar, R. Ananthanarayanan and D. Saha, "Testing Framework for Black-box AI Models," *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, 2021, pp. 81-84, doi: 10.1109/ICSE-Companion52605.2021.00041. **Video**: https://youtu.be/984UCU17YZI
  - Machine Learning Testing: Survey, Landscapes and Horizons, Jie M. Zhang, Mark Harman, Lei Ma, Yang Liu, https://arxiv.org/abs/1906.10742, 2019
  - Software Engineering for AI-Based Systems: A Survey, Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, Stefan Wagner, https://arxiv.org/abs/2105.01984, 2021

What is AI Being Tested For?

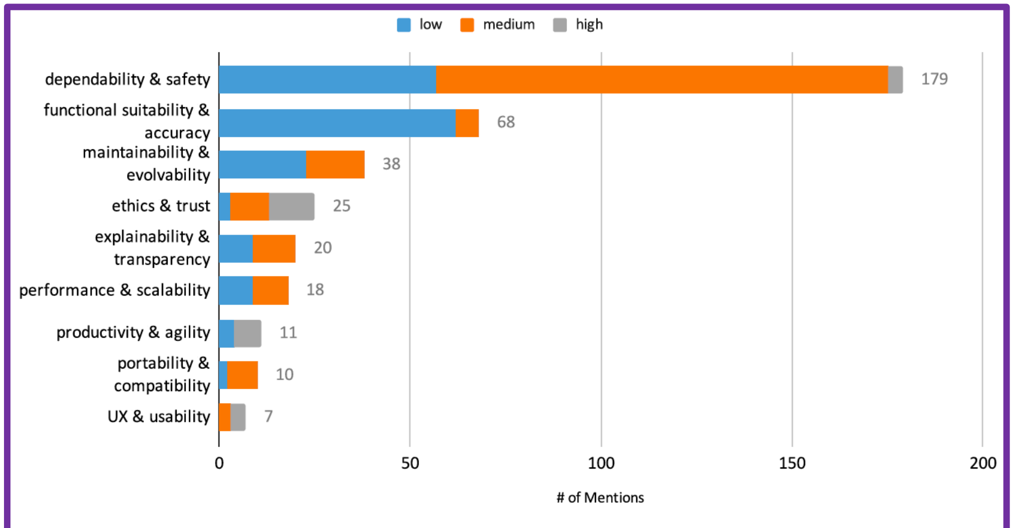


Fig. 16. Frequency of quality goals per thematic cluster and level of abstraction.

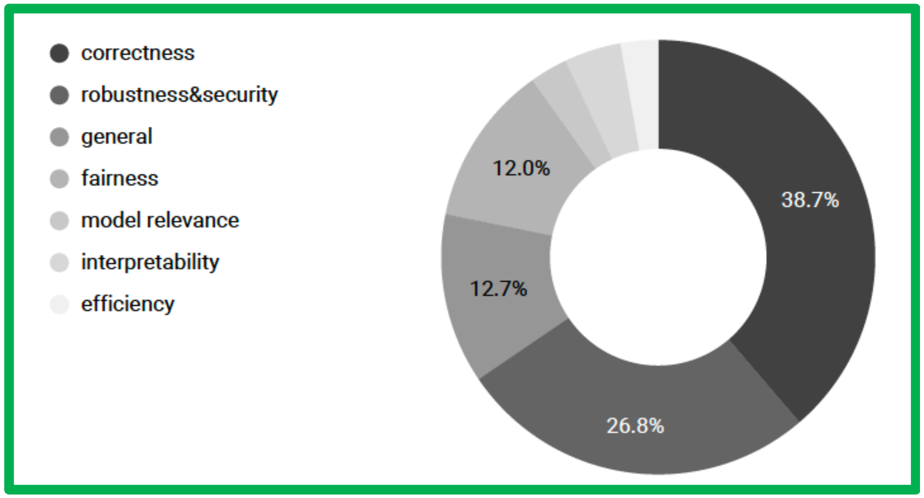


Figure Source: Software Engineering for AI-Based Systems: A Survey, Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, Stefan Wagner, https://arxiv.org/abs/2105.01984, 2021

Figure Source: Machine Learning Testing: Survey, Landscapes and Horizons, Jie M. Zhang, Mark Harman, Lei Ma, Yang Liu, https://arxiv.org/abs/1906.10742, 2019

Figure Source: Machine Learning Testing: Survey, Landscapes and Horizons, Jie M. Zhang, Mark Harman, Lei Ma, Yang Liu, https://arxiv.org/abs/1906.10742, 2019

ML Testing

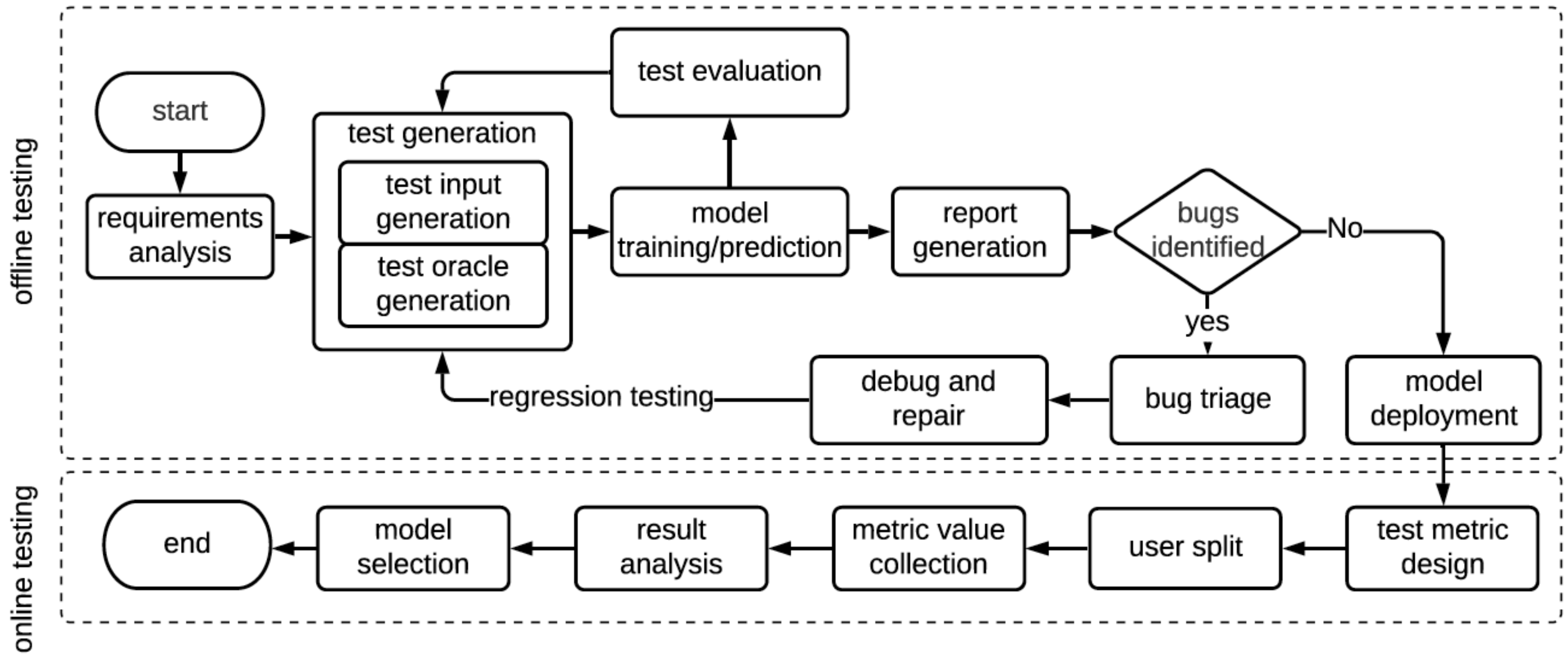


Figure 5: Idealised Workflow of ML testing

Comparison of Testing Needs

Table 1: Comparison between Traditional Software Testing and ML Testing

Characteristics	Traditional Testing	ML Testing
Component to test	code	data and code (learning program, framework)
Behaviour under test	usually fixed	change overtime
Test input	input data	data or code
Test oracle	defined by developers	defined by developers and labelling companies
Adequacy criteria	coverage/mutation score	unknown
False positives in bugs	rare	prevalent
Tester	developer	data scientist, algorithm designer, developer

Figure Source: Machine Learning Testing: Survey, Landscapes and Horizons, Jie M. Zhang, Mark Harman, Lei Ma, Yang Liu, https://arxiv.org/abs/1906.10742, 2019

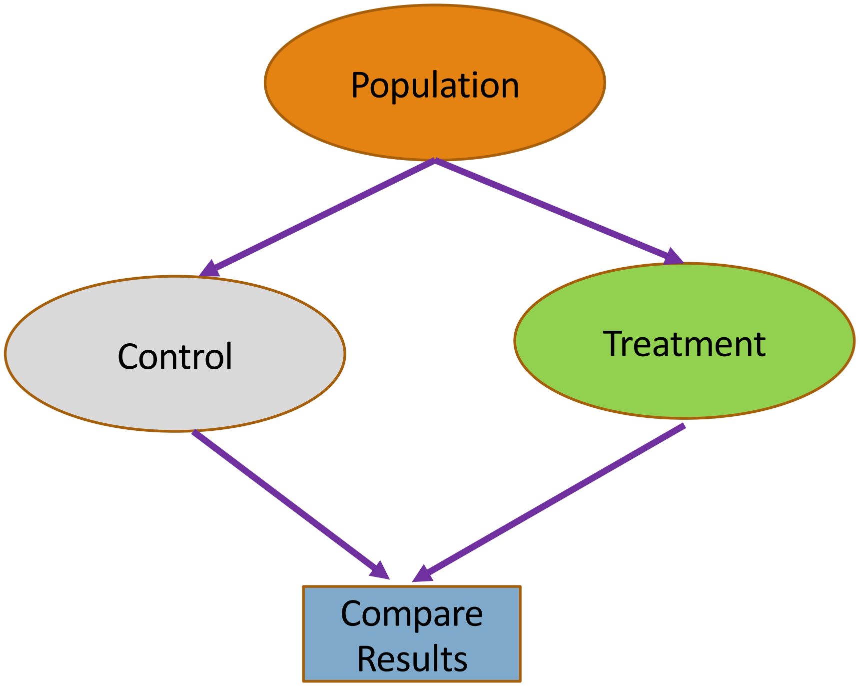
Causal Evaluation

Randomized Control Trial

- The population is ***randomly*** divided in two groups
- Control group gets placebo (nothing changes)
- Treatment group gets actual benefit that is being tested
- The difference in outcomes from the two groups are checked for statistical significance

Notes:

- population has to be large
- the split has to be random
- difference in treatment has to be ethical
- Considered gold standard in testing in critical domains like medicine, public policies



Example

Impact of mask on COVID cases

- https://theconversation.com/a-new-data-driven-model-shows-that-wearing-masks-saves-lives-and-the-earlier-you-start-the-better-149621

RCT for AI-based Systems

- Not very common for AI but increasing; against the prevalent culture
- Takes more effort than technical or business evaluation

Concluding Segment

Lecture 24: Concluding Comments

- We looked at how to evaluate explanation methods
- AI and testing has two connotations
  - How AI has been used for software testing ?
  - How software testing for AI has to be done?
- We looked at current testing practices
- There is need for randomized control trial to build lasting trust in AI

About Next Lecture – Lecture 25

Schedule Snapshot

Oct 26 (Tu)	**Review: Explanation Methods, AIX 360, Discussion**	Quiz 3
Oct 28 (Th)	**Review: project presentations, Discussion**	
Nov 2 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	
Nov 4 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods	
Nov 9 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods	
Nov 11 (Th)	Explanation Methods Trust: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	Quiz 4

Lecture 25:

- Human AI Collaboration
- Chatbots
- Trust issues with Chatbots