AI Ethics, Explainability & Regulations

Executive Education Program in Artificial Intelligence and Machine Learning – Batch-3

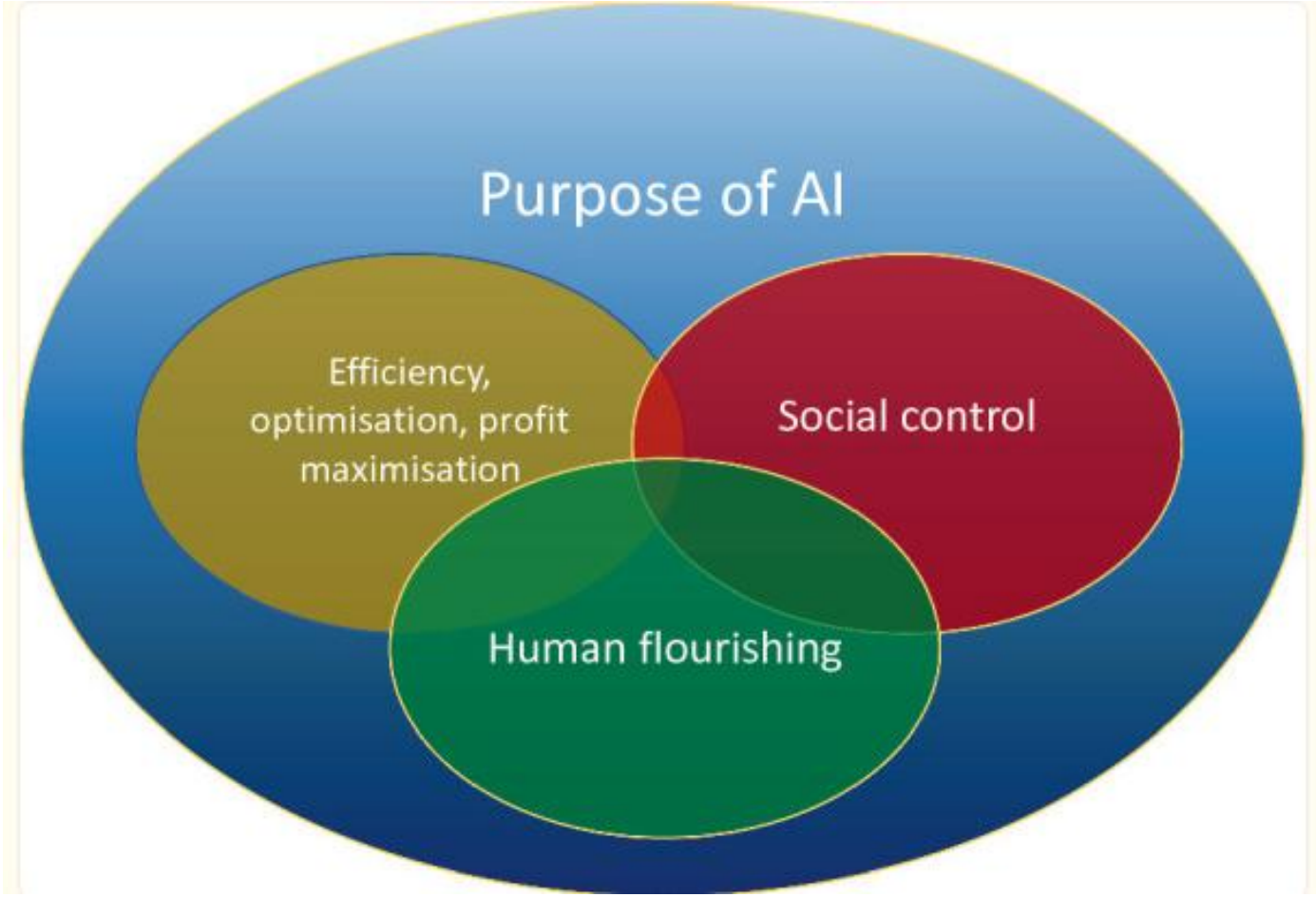
IIM Visakhapatnam

April 12, 2025

What is AI?

- Definition:
  - *The ability of a computer or a computer-enabled robotic system to process information and produce outcomes in a manner similar to that of the thought process of humans in learning, decision making and problem solving*
- How does it usually work
  - Machines are fed with large amount of training data
  - This data is labeled/unlabeled
    - The machine tries to identify patterns in the input data and tries to predict the outcome for unseen cases – Supervised learning
    - The machine tries to identify patterns in 'unlabeled' data without any information on input or output – Unsupervised learning
- Why is it widespread?
  - Miniaturization of computing power, Networking of sensors and devices, Affordable internet access

What is AI Used For?



Why shall we discuss AI Ethics?

AI Usecases

- Movie studios predicting which scripts are likely to succeed
- Banks predicting who is likely to default on a loan
- Governments predicting who will become radicalized
- Governments predicting who is a terrorist
- Judges predicting who won't (re)offend when out on bail
- Google predicting to whom to display ads for high-paying executive jobs
- Employers predicting which hires will be the best
- Police predicting where crime will be
- Governments predicting which cops will commit misconduct
- Autonomous underground drilling machine
- Autonomous cars

AI & Ethics – Corporate View

- Google
  - *"New products and services, including those that incorporate or utilize artificial intelligence and machine learning,* ***can raise new or exacerbate existing ethical, technological, legal, and other challenges,*** *which may negatively affect our brands and demand for our products and services and adversely affect our revenues and operating results."*
- Microsoft
  - *"AI algorithms may be flawed. Datasets may be insufficient or contain biased information.* ***Inappropriate or controversial data practices*** *by Microsoft or others could impair the acceptance of AI solutions. These deficiencies could undermine the decisions, predictions, or analysis AI applications produce, subjecting us to* ***competitive harm, legal liability, and brand or reputational harm.***"

Philosophy of Ethics

- What is Ethics?
  - *Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues*
- Tech-centric focus:
  - Solely revolves around improving the capabilities of an intelligent system
  - Does not sufficiently consider human needs and societal values and ethics
- Ethical AI – designed and developed in a manner that is aligned with the values and ethical principles of society and community it affects
- Deontological (action itself) vs. Utilitarianism (action based on consequences) view
- Other views: Virtue ethics; feminist ethics of care; religion-based ethics

Philosophy of Ethics

- Deontological view
  - Does an action follow a moral rule?
  - End does not justify the means - Wrong to lie, steal or murder under all conditions
  - Too strict? Lying to a soldier in Nazi Germany
- Utilitarianism view
  - Does an action produce net good consequences?
  - Is end irrelevant?
  - Must accept that lying, stealing or murder could in theory be ethical if the net benefit of the action results in greater cumulative well being

AI & Ethics - Thought Experiments

- https://www.youtube.com/watch?v=ixIoDYVfKAo&feature=youtu.be
  - Driverless Cars & Ethics – Reaction Vs. Decision

- https://www.youtube.com/watch?v=yg16u_bzjPE – Trolley problem

- Dept. of Education trying to decide whether to let a computer predict children who might be at risk of falling behind at school
  - How does it fare as per both approaches?
- Other Examples: Alexa, Target

Isaac Asimov's "Three Laws of Robotics"

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Is such an AI system possible??

Sources of Ethical Risk in AI

- Privacy
  - Individual or Collective
  - Privacy by Design: Seven Principles
    - Proactive, not reactive; Privacy as default setting; End-to-end security etc.

- Bias
  - In Data, Method, and Mind
  - Tone deaf to Diversity and Inclusion, i.e., discrimination
  - Biased assumptions in developing, understanding, and deploying models

- Explainability & Interpretability in AI algorithms

Data Privacy: Individual & Collective

Personal Data Privacy

- Personal data classification
- Reasonable expectation – consumer awareness
- Consent Framework - boilerplate
- Obligation of Data Fiduciaries

What is Personal Data?

- Personal Data – whether the data is related to an identified or identifiable individual
  - Protection of personal data ~ Objective of protecting an individual's identity
  - Developments in data science have changed the understanding of identifiability
- Methods of Removing Identification
  - Anonymization
    - Process of removing identifiers from personal data in a manner ensuring that the risk of identification is negligible
    - Mathematical and technical methods to distort data irreversibly ensure that identification is not possible
  - Pseudonymization
    - Method by which personal identifiers are replaced with pseudonyms
    - Important component of privacy by design
    - Carries a risk of re-identification without specific technical and organizational measures
  - Anonymized Data (no more personal data) vs. Psuedonymized Data (personal data)

Sensitive Personal Data

- Data integral to an individual's identity
- Processing of this data can lead to graver concern, hence stricter rules
- Criteria to 'categorize' data as sensitive:
  - Likelihood of causing significant harm to the individual
  - Expectation of confidentiality to that category of data
  - Significant discernible class of data principals could suffer harm or a similar nature
- Residuary power vested with the DPA (Data Protection Authority)
- *Passwords, Financial data; health Data; Official identifiers including government-issued identity cards; sex life and sexual orientation; biometric & genetic data; transgender status or intersex status; caste or tribe; and religious or political beliefs or affiliations*

Consent

- *Expression of a person's autonomy or control, consequent to allowing another person to legally disclaim liability for acts which have been consented to*
- Enabled through notice – affirmative obligation placed on data fiduciaries
- Advantage of consent – in principle
  - It respects user autonomy
  - Provides a clear basis for the entity to whom the consent is given to disclaim liability, if reqd.
  - Hence, the meaningfulness of consent shall be carefully determined
- Current Operational Framework
  - Complex and boilerplate
  - Unequal bargaining power of parties and ineffective in informed consent

New Consent Framework

- Informed Consent
- Explicit Informed Consent

Predictive Analytics: Collective Data Privacy

- In human society context:
  - Leverages large behavioural data sets to classify individuals according to future risks, economic developments or expected costs, risk, utility functions
- What is the prediction used for?
  - Sensitive attributes, future behaviour, risk/utility functions – insurance cost, criminal recidivism
- Goal of PA – Discrimination! – In a positive sense
- Can be for leveraged for classification or predicting continuous values

Predictive Analytics Ethics

- Some predictions that PA can make
  - Whether you will quit your job; Whether you are likely to die soon
  - Which race you belong to – e.g., facial recognition to identify and track Uighurs – first known case of government using ML to profile by ethnicity
  - Short-list job applicants; Criminal recidivism scoring
  - Individual risk score - Differential insurance premium pricing
- Is there anything wrong about this?
- Does this come under mishandling/leaking/stealing data?
- If a model predicts incorrectly, there's a cost involved; what if it predicts too correctly?

What is Predictive Privacy?

- Predictive privacy
  - When sensitive information about that person/group is predicted against their will or without their knowledge based on data of many other individuals, leading to decisions that impact anyone's social/ economic, psychological, physical well-being or freedom
- Conditions?
  - Not by stealing/leaking information about a specific person, but by deriving a prediction based on others' data
  - Proxy data of the individual for whom prediction is made may have been acquired lawfully
  - Prediction need not be accurate for a violation to occur
- To protect individuals/groups against unfair treatment and infringements on their autonomy, dignity, and well-being, using information predicted by leveraging statistical correlations with others' behaviour

Solution

- Regulations?
  - Personal Data Protection Laws
- Technology?
  - Ethical evaluation of PA systems – at every stage
  - Model verification/Feedback loops
  - Privacy by design

Model Verification/Feedback Loop

- Imbalance in detecting mispredictions between false positives and false negatives
- Hiring Algorithms
  - No data on false negatives: Keeping track of rejected applicants is hard
  - More incentive to remove false positive – selected but not performing well
- Criminal recidivism scoring
  - False negatives detection easy – outside, but may re-offend
  - False positives – In prison considering they may re-offend – hard to detect
- Credit scoring

AI/ML Bias/Algorithmic Bias

- Occurrence of biased results due to human biases that skew the original training data or AI algorithm – leading to distorted outputs and potentially harmful outcomes
- Example: Gorrilla in GooglePhotos
- Quote:
  - *"The world according to Stable Diffusion is run by white male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes, while women with dark skin flip burgers."*
- Types of Bias
  - Availability Bias
  - Ingroup/Outgroup Bias
  - Sunk Cost Bias
  - Stereotyping
  - Confirmation Bias
  - Self-serving Bias

Examples of ML Bias

- Bias in design – the beliefs of engineers
- Bias in how data is collected, encoded and published for AI can be biased
  - Dermatologists always use a standard ruler to measure the size of the lesion. If it is greater than 3 cm, they choose to do a biopsy. Neural Network assumed every picture with a ruler was malignant
  - Black patients assigned the same level of risk by the algorithm are sicker than White patients – Why?
    - Racial bias reduces the number of Black patients identified for extra care by more than half.
- Here are the victims. Where is the perpetrator?

AI Explainability

- Setting where the models are being used
  - High-stakes decision making - Impact on human life, health, finances
  - Recommending movies, books, friends etc.
- Do all of these require model understanding? Why?
  - Little or no consequences for making incorrect predictions
  - Problems are well-studied and extensively validated – we can trust model predictions
- What combination requires model accuracy and explainability?
  - High-stakes decision making - Impact on human life, health, finances
  - Problems are not well-studied and extensively validated
  - Accuracy is not enough

How Model Understanding Helps?

- Prediction of Siberian Husky
  - Model is focussing on snow rather than the animal
  - Model understanding facilitates debugging
- Whether people should be released on bail or not? (Person too risky to release)
  - Model understanding facilitates bias detection
- Denied loan
  - Model understanding offers actionable information to individuals who are adversely affected by model predictions
- Doctor is detecting a disease
  - Model understanding helps assess if and when to trust model predictions when making decisions – for vetting the models – can it be approved for broader deployment

Why Model Understanding?

Utility

- Debugging
- Bias Detection
- Recourse
- If and when to trust model predictions
- Vet models to assess suitability for deployment

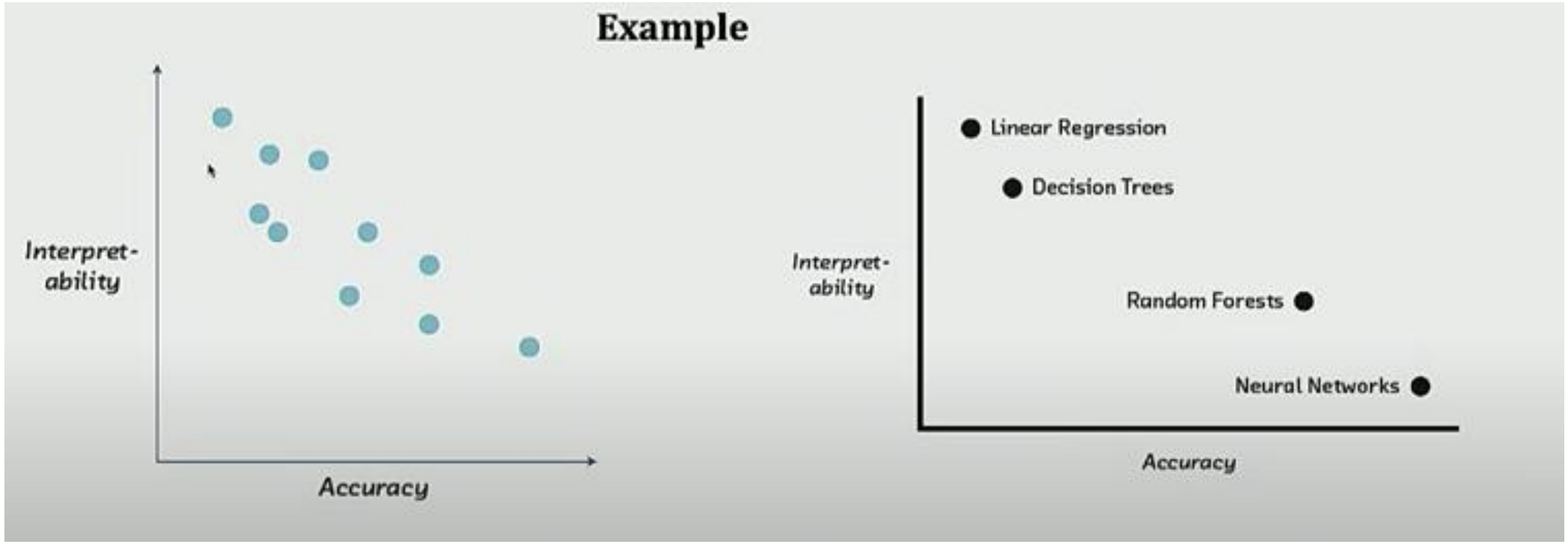
Stakeholders

- End Users (e.g., Loan applicants)
- Decision makers (Doctors Judges)
- Regulatory agencies (FDA)
- Researchers and Engineers

Achieving Model Understanding

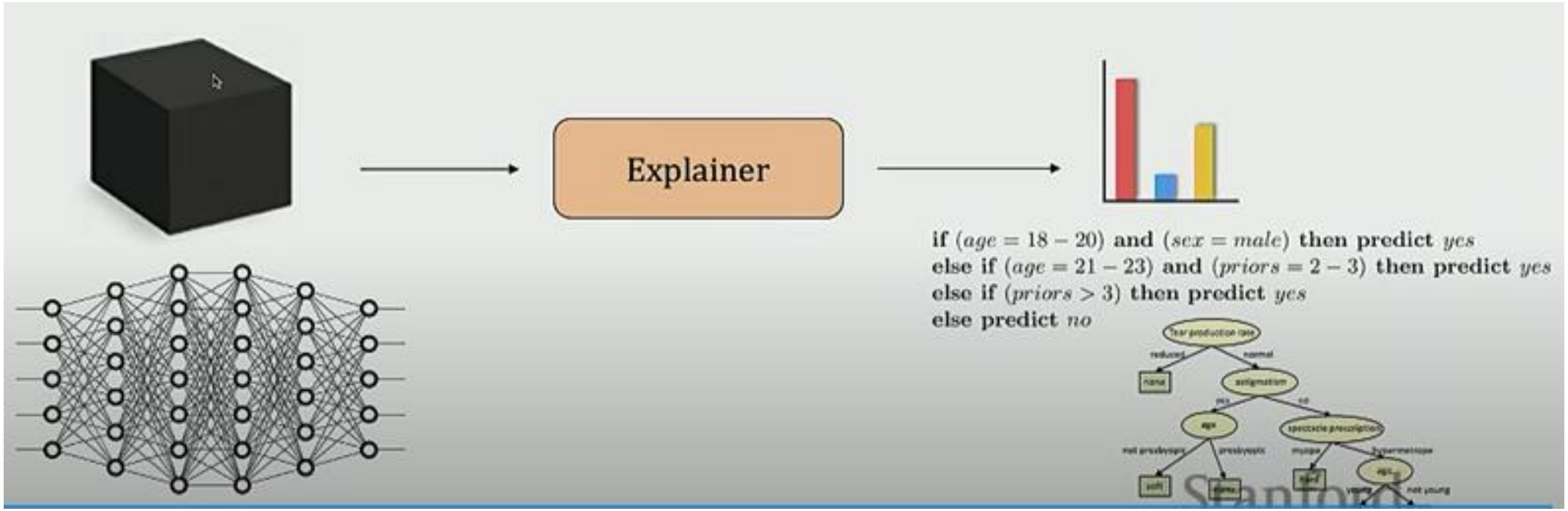
- Build inherently interpretable predictive models
  - Linear Regression, Logistic regression, Shallow decision trees, Small rule-based models
- Explain pre-build models in a post-hoc fashion
  - Black-box model – pass them through explainer – output gives important features on which the decision is being made
  - Black-box and white-box approaches
- Accuracy versus Interpretability

How do decide between the two?



Source: Stanford Seminar – ML Explainability Workshop

Achieving Model Understanding



Source: Stanford Seminar – ML Explainability Workshop

Categorization of XAI Methods

- Model agnosticity
  - Model-agnostic & Model-specific
- Scope of provided explanations
  - Global explanation & Local explanation
- Data Types
  - Graph, Image, Text/Speech, Tabular
- Explanation Type
  - Visual (correlation plots), Data points, Feature Importance, Surrogate models (simple models in lieu of complex ones)

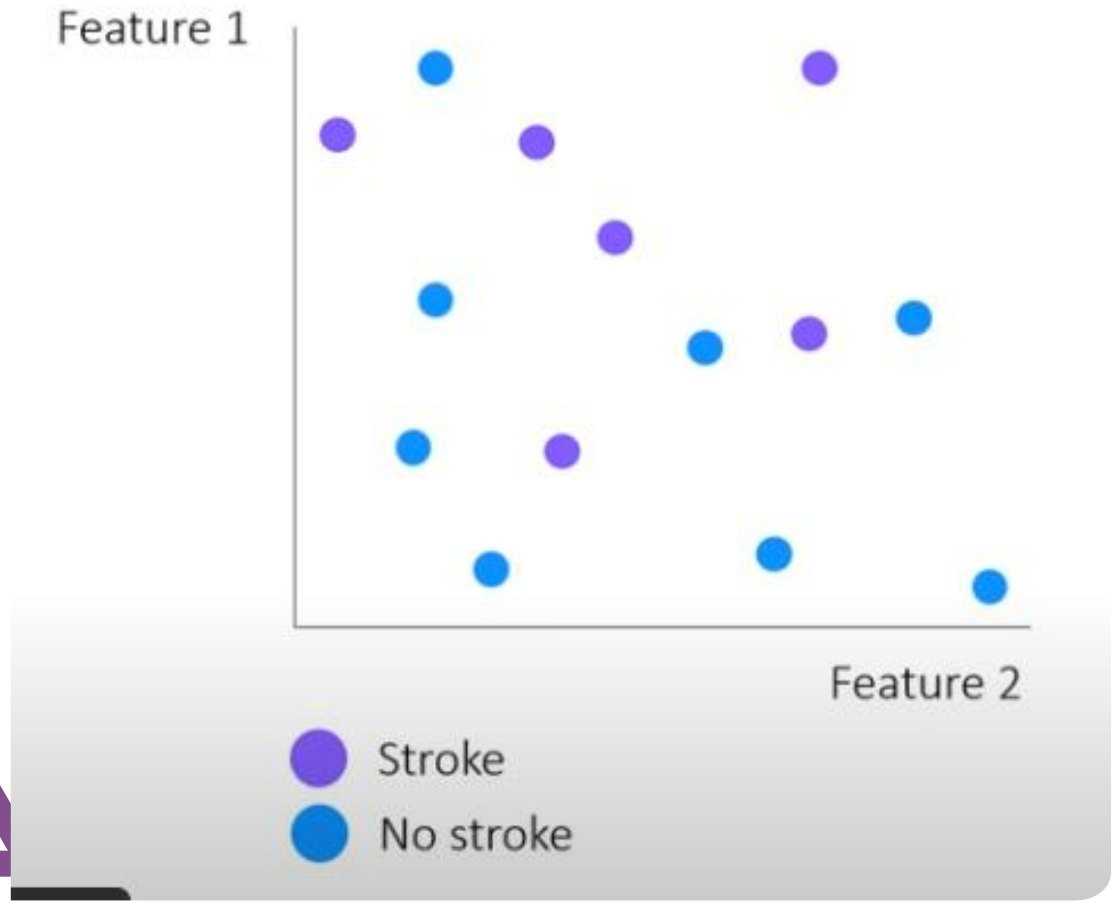
By-design Interpretable Models

- Glass-box models
  - Linear Regression
  - Decision Tree
  - Logistic Regression
  - Explainable Boosting Machine
- Predict for a particular person
  - For example: Whether someone will get a stroke or not
- How to interpret each of the above glass-box models?
  - Interpret library
  - Linear regression
    - Lr.explain_local – for any specific datapoint or a collection of them
    - Lr.explain_global – for the model

By-design Interpretable Models

- Explaining Decision trees
  - Entropy, information gain, tree-depth
  - Explanation – a set of if-else statements; depth can be a problem
  - Tree.explain_local; tree.explain_global

LIME: Local Interpretable model-agnostic explanations



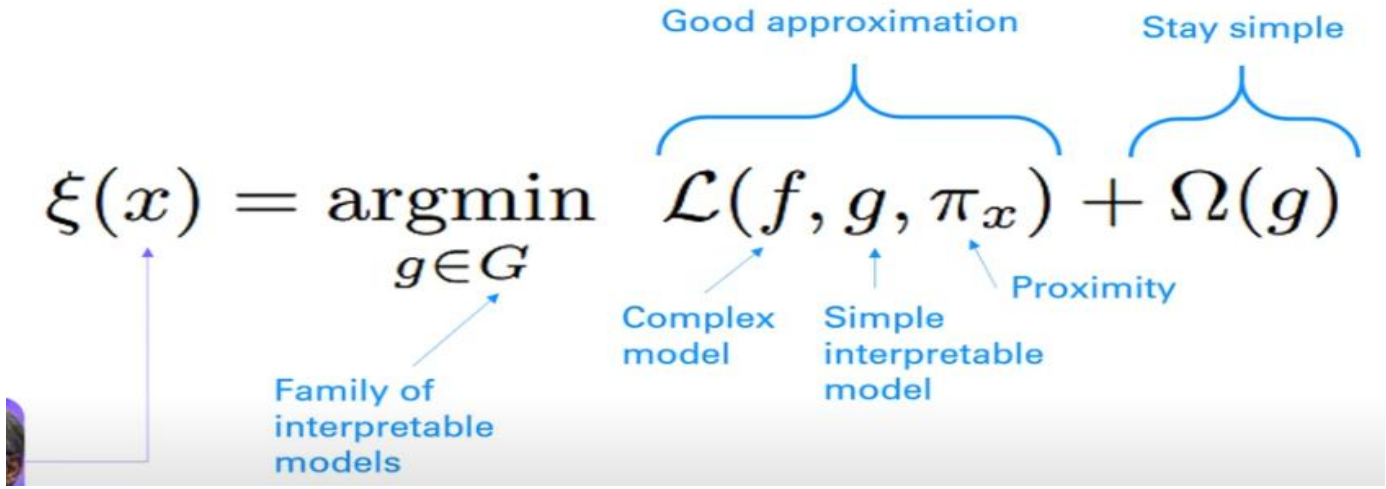
- Non-linear relationship
- Difficult to explain relationship (whole decision boundary into one explanation)
- Move from global to local explanation
- Relative impact of features on each decision – create a surrogate model

LIME

- Works on any black box model
- Works only on the inputs and outputs of the model – internal working of the model is hidden
- Any kind of data types
- Explanations can be validated leading to trust in the module
- Explanations are locally faithful but not necessarily globally

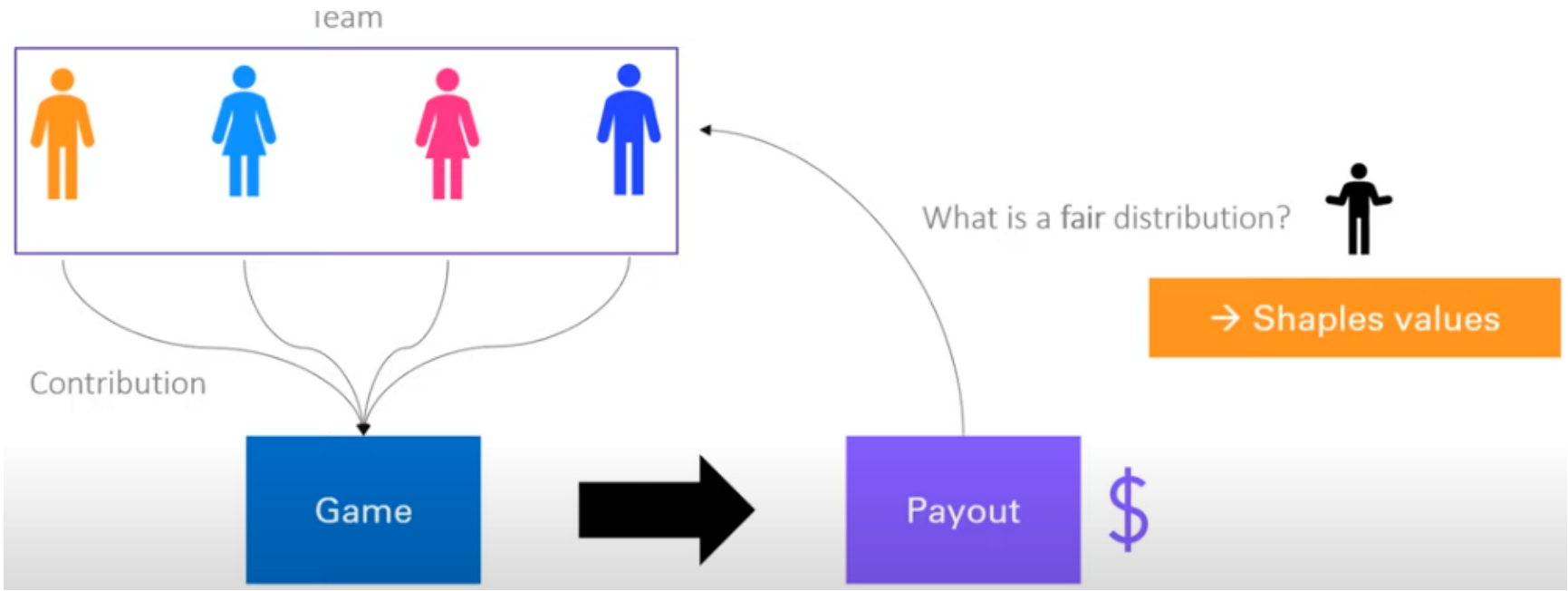
Lime

- The Math in Lime



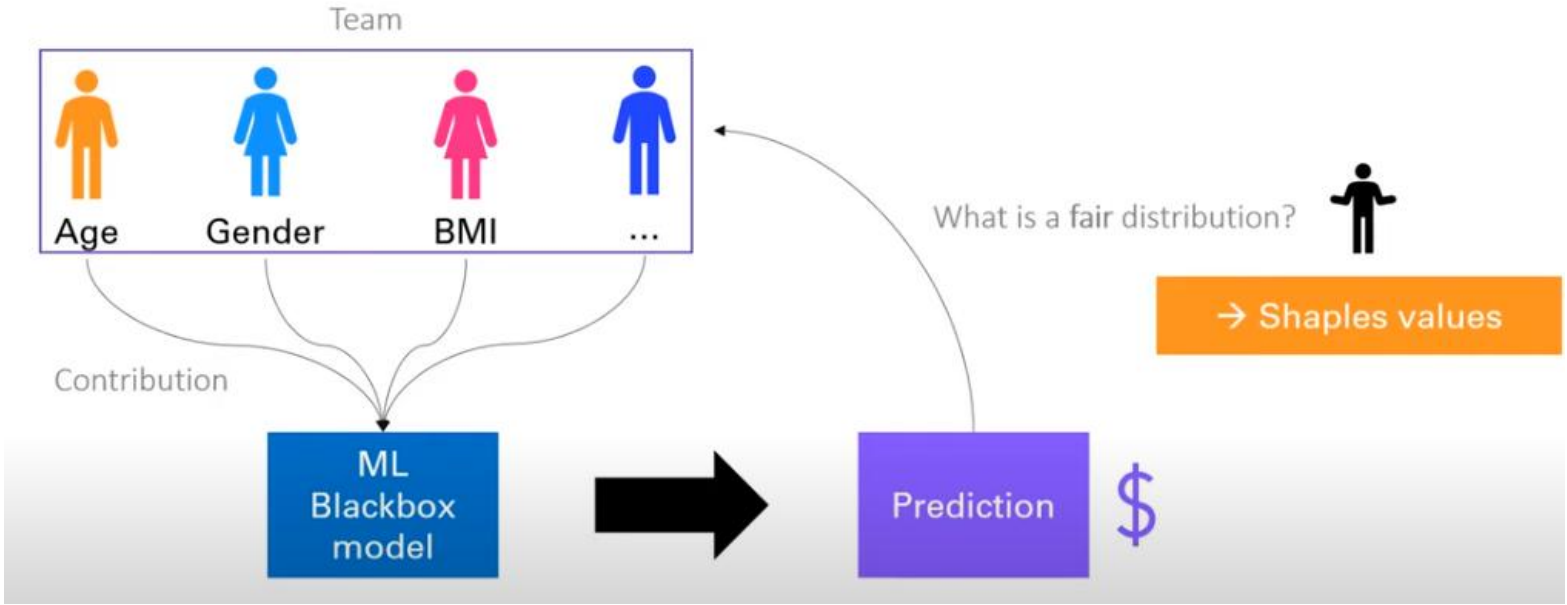
SHAP: Shapley Additive Explanations

- Cooperative Game Theory
- How does the game outcome change with or without a particular person



SHAP: Shapley Additive Explanations

- Primarily Local explanations – can be aggregated for global explanations



How to decide between the two?

- If you can build an interpretable model which is fairly adequate for your setting, go ahead and do it
- Else, post hoc explanations come to the rescue
- Problems with Post hoc explanations
  - Considers correlates instead of actual variables
- Interpretability vs. Explainability

Tools to make AI Ethical

- Aequitas – helps measure bias in uploaded data sets
  - Open-source bias audit toolkit for ML developers, analysts and policymakers to audit ML models for discriminations and bias
  - http://aequitas.dssg.io/example.html#audit-results-bias-metrics-values
  - Bias metrics considered: demographic disparity, impact disparity
  - http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/
  - Detection of probability of false positives & false negatives
- CompTIA
  - advises ensuring balanced label representation in training data
  - makes sure the purpose and goals of the AI models are clear enough so that proper test datasets can be created to test the models for biases.
- Python Libraries
  - Themis-ml - reduce data bias using bias-mitigation algorithms.
  - HolisticAI library
  - BiasDetector Package - https://pypi.org/project/bias-detector/
  - Checks for Statistical Parity, Equal Opportunity, and Predictive Equality

Tools to make AI Ethical & Explainable

- InterpretML
  - Model debugging - Why did my model make this mistake?
  - Detecting fairness issues — Does my model discriminate?
  - Human-AI cooperation — How can I understand and trust the model's decisions?
  - Regulatory compliance — Does my model satisfy legal requirements?
  - Focus on high-risk applications — Healthcare, finance, judicial
  - Functionality
    - Explore model attributes such as performance, global and local features and compare multiple models simultaneously. Run what-if analysis as you manipulate data and view the impact on the model
    - Both Explainable and deep neural networks
    - Analyzes relation between input features and output predictions to interpret models
- GitHub Project
  - Awesome-machine-learning-interpretability project

AI Regulations: EU AI Act

- Human-centric and ethical development of AI in Europe
- AI Classification – on the basis of risk
  - Minimal Risk – AI-enabled video games, Spam filters – Not much to be done
  - Limited Risk – Deep fakes – Be transparent
  - High Risk – AI in critical areas – Profound impact – Self-driven cars, Healthcare – Fairness is critical
    - Thorough risk assessment, Top-quality data, Maintain detailed logs, Human supervision
  - Unacceptable Risk – Social scoring – Banned in EU
- Stifling Innovation - ?
- Status in India - ?

Five Areas of Ethical Focus for AI (IBM)

- **Accountability:** Designers & developers are responsible for AI design, development & outcomes
- **Value Alignment:** Design to align with the norms and values of your user group in mind
- **Explainability:** Design for humans to easily perceive, detect, and understand its decision process
- **Fairness:** Design to minimize bias and promote inclusive representation
- **User Data Rights:** Design to protect user data and preserve the user's power over its access & uses

Ethically Aligned Design: Three Pillars (IEEE)

Universal Human Values	• Designed to respect human rights, align with human values • Holistically increase well-being while empowering as many people as possible
Political Self-Determination & Data Agency	• Designed to nurture political freedom and democracy with people having access and control over their data • People have agency over their digital identity
Technical Dependability	• Reliably, safely, and actively accomplish the objectives for which they were designed • Validation and verification processes shall be designed

Ethically Aligned Design: General Principles (IEEE)

- Human Rights - to respect, promote, and protect internationally recognized human rights
- Awareness of Misuse – shall guard against all potential misuses and risks
- Well-being - human well-being as a primary success criterion for development.
- Data Agency - maintain people's capacity to have control over their identity.
- Effectiveness - provide evidence of the effectiveness and fitness for purpose
- Transparency & Accountability – Basis of decision shall be discoverable; Unambiguous rationale
- Competence - creators shall specify & operators shall adhere to the knowledge required for effective operation

DEEP-MAX Scorecard: TN's AI Policy

- Diversity
  - Trained for diversity in race, gender, religion, color, features, food habits etc.
- Equity & Fairness
  - Does the system promote equity and treats everyone fairly?
- Ethics
  - Preserves human values of dignity, fairness, respect, compassion and kindness for a fellow human being
- Privacy & Data Protection
  - Privacy and data protection features built-in?
- Misuse protection
  - Features that inhibit or discourage the possible misuse?
- Audit & Transparency
  - How good is auditability of decisions made by the autonomous system?
- Cross Geography & Society
  - Across geographies and societies, especially for the disadvantaged sections

Why is it hard to implement in practice?

- Companies do not disclose what algorithms they use or the data they use to train them?
  - Intellectual Property issues
  - To prevent a security breach
  - Complexity issue too – vendor uses a combination of public and private datasets
- Problems in understanding the logic of a statistical model
  - Not as simple as a linear regression
  - Trade-off
    - Complex algorithms unlock capabilities simpler statistical models cannot handle – but at the cost of explainability
    - Simpler models may be easier to explain, but have biases and assumptions that influence what they see in the data
  - Difficult to know what the model optimizes for – One business process may require different ML models with varying requirements of explainability
    - Offering credit card, the two processes are: evaluate risk and approve the card (requires greater explainability) and to predict propensity to convert and personalize the offers (lower explainability)

Some Quotes to Motivate Further

- *We kill people based on metadata*
  - General Michael Hayden, (former director, NSA and CIA)

- *Technology has always been a double-edged sword. Fire kept us warm and, cooked our food and burned down our houses.*
  - Ray Kurzweil

References

- https://www.onetrust.com/blog/principles-of-privacy-by-design/
- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G - Amazon scraps secret AI recruiting tool that showed bias against women
- https://www.science.org/doi/10.1126/science.aax2342 - Dissecting racial bias in an algorithm used to manage the health of populations
- https://policyoptions.irpp.org/magazines/march-2018/ai-in-government-for-whom-by-whom/ - AI in government: for whom, by whom?
- https://policyoptions.irpp.org/magazines/march-2018/ai-automating-inequality/ - Will AI just wind up automating inequality?
- https://medium.com/nerd-for-tech/an-brief-overview-of-some-ethical-ai-toolkits-712afe9f3b3a

References

- Interpretable Machine Learning: A Guide for making Black Box Models Explainable – by Christoph Molnar
- https://www.youtube.com/watch?v=OZJ1IgSgP9E&list=PLV8yxwGOxvvovp-j6ztxhF3QcKXT6vORU