Artificial Intelligence & Machine Learning in Clinical Research and More

Kao-Tai Tsai
Frontier Informatics Services
(kaotai.tsai@gmail.com)

November 2, 2026

Outline of Presentation

1. The good old days
2. What is Machine Learning
3. CRAN Task View: ML & Stat Learning
4. Generative AI (GAI)
5. Classical Stat Data Analysis vs. GAI era ML
6. Potential resistance to adapt AI/ML
7. Needs to explain outputs of ML
8. Scopes to Model the Explanation Functions
9. Methods to model the Explanation Functions
10. SHAP (SHapley Additive exPlanations)
11. A clinical study example
12. Applications of AI/ML on clinical research
13. Applications of AI/ML on medical devices
14. Summary

Statistical Computing by Fisher

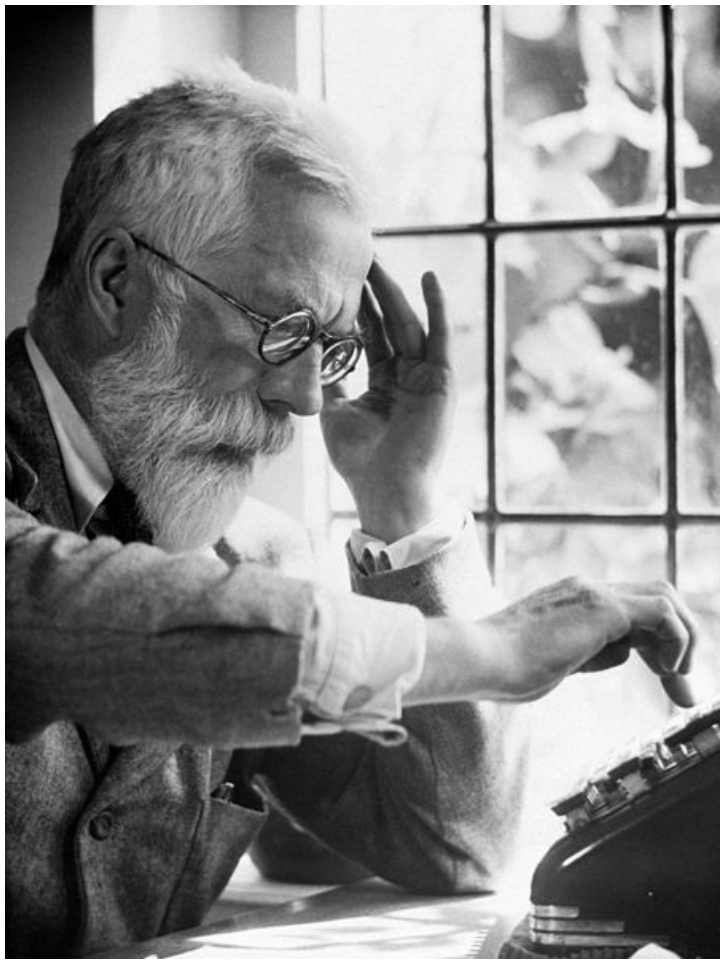


Figure 1: Statistical Computing: Fisher in 1954 at Gonville & Caius College, Cambridge.

The evolving spiral path of statistics [6]

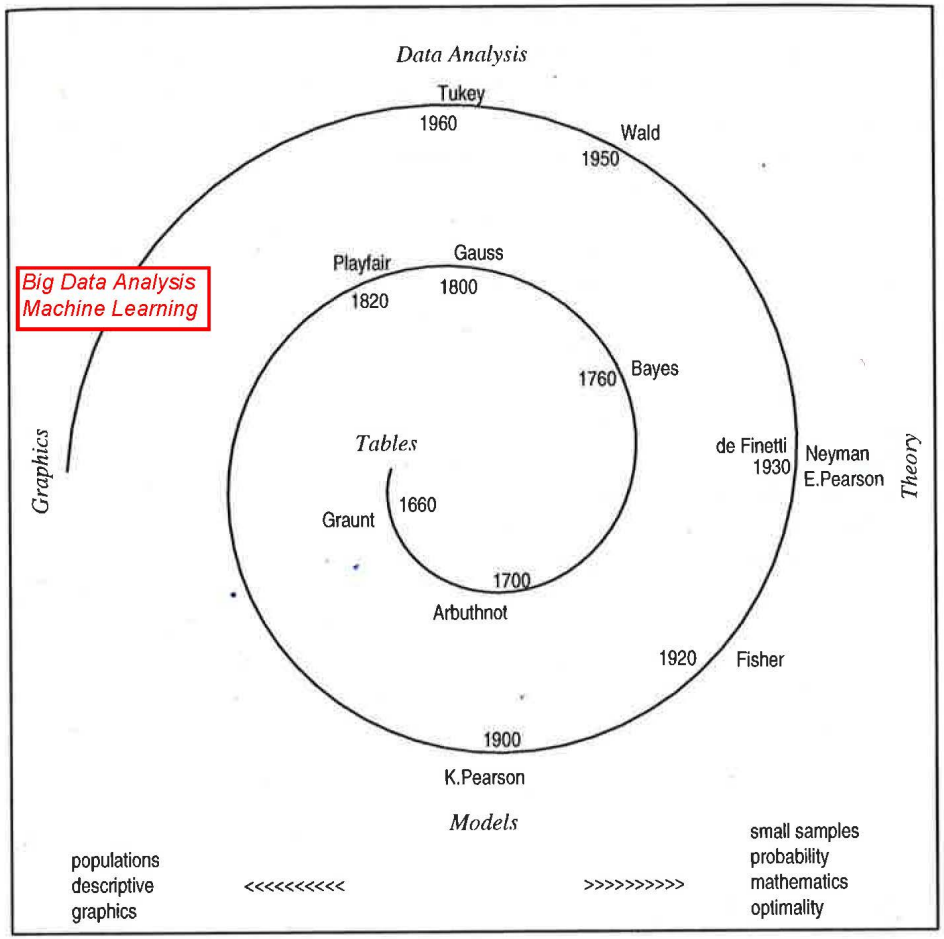


Figure 2: The evolving spiral path of statistics (Huber 1997)

The concept of Precision Medicine - 1

Thrust for Big Data Analysis and Machine Learning

- In 2015 US President Obama proposed the initiative of **Precision Medicine** during his State of the Union address.
- Following this there was a huge increase in interest and research on precision medicine methodologies.

Big Data sources and analytics

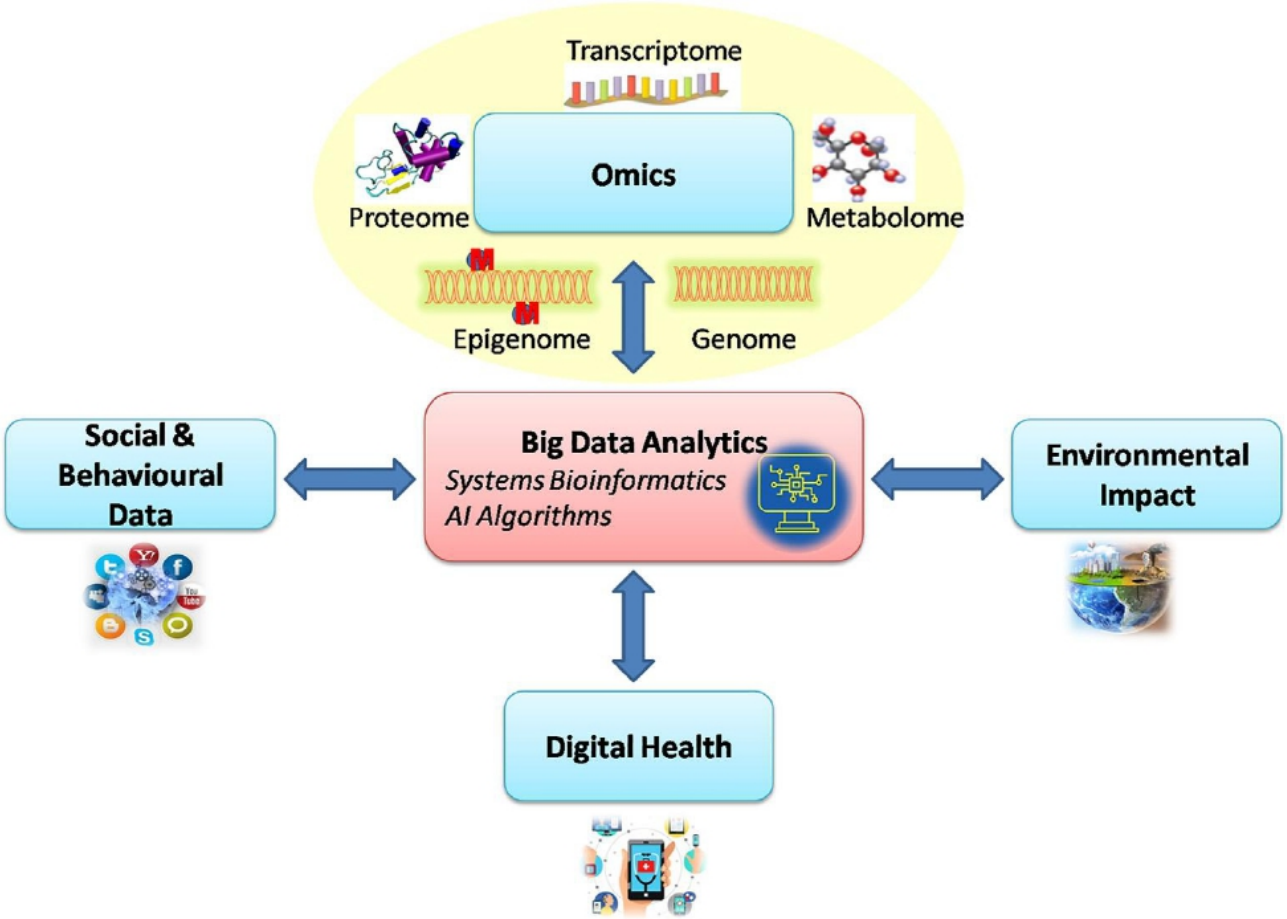


Figure 3: Big Data sources and analytics

Big Data sources and analytics: challenges

But: the facts!

- Varieties of data types: numeric, text, images, etc.
- Data are not all from controlled design experiments.
- How to analyze and pull all the information from various sources to better understand the sciences.
- AI/ML becomes the natural choice of tools to consider.

A framework of AI life cycle

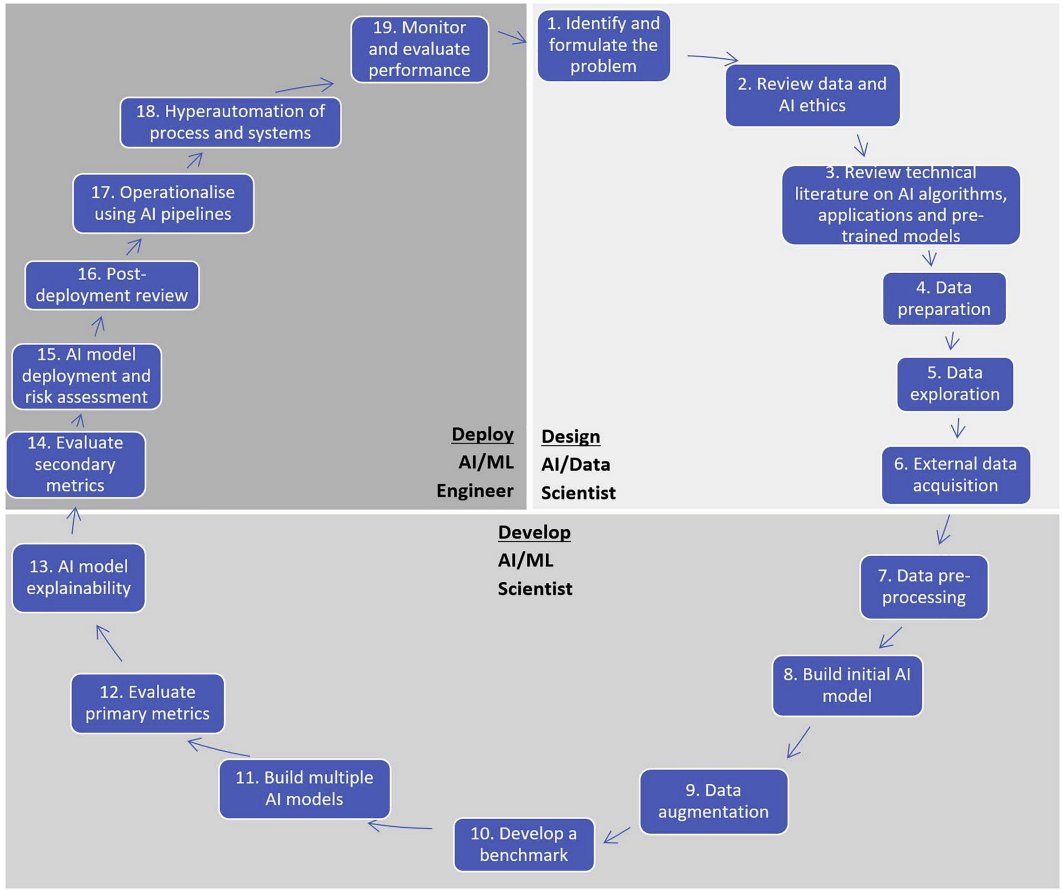


Figure 4: The CDAC AI life cycle: three phases: design, develop, and deploy

What is Machine Learning - 1

- Machine learning (ML) is a field of study in AI concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without being explicitly programmed.
- Advances in the field of deep learning (DL) have allowed neural networks, a class of statistical algorithms, to surpass many previous ML approaches in performance.
- Statistics and mathematical optimization methods compose the foundations of ML.
  (»> analytic)

What is Machine Learning - 2

- From a theoretical viewpoint, probably approximately correct learning provides a mathematical and statistical framework for describing ML.
- Most traditional ML and DL algorithms can be described as empirical risk minimization under this framework.

What is Machine Learning - 3

Definition of ML

- Tom M. Mitchell [8] provided a widely quoted, more formal definition of the algorithms studied in the ML field:
  " A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E ."
- This definition of the tasks in which ML is concerned is fundamentally operational rather than defining the field in cognitive terms.

What is Machine Learning - 4

Modeling with ML techniques

- Any valid data analysis method can be used to build a ML model;
- the common purposes of data analysis include prediction, classification, clustering, optimization, association, neural networks and deep learning, visualization, etc.
- The data can be cross-sectional or longitudinal structures in nature, it can also include multiple data sets measured on various aspects of the study subjects.
- The computational procedures are usually based on certain pre-defined optimization criteria.
- Most of the methods can be found in the R-software package [3] or in Python software packages.

What is Machine Learning - 5

- It is rare to have one single model which serves a good purpose for different sub-populations in the whole data set, this is especially true when dealing with big size data,
  therefore, it is important to divide the original dataset into several more homogeneous datasets, and analyze these datasets individually.
- A useful method to search for the more homogeneous subsets is recursive partition based on the input data variables.
  This method was introduced by the authors in [1] and have been widely extended into several popular methods, such as `randomForest`, etc.

What is Machine Learning - 6

- It is recommended practice to begin the model development process with a simple algorithm, default parameters, and default architecture.
  This enables a rigorous and validated approach that can also be interpreted, presented, and documented.
- The model should be evaluated using the metrics designed for the type of AI capability and the algorithm itself.
- This initial model serves several purposes.
  Firstly, if the intended AI approach is feasible, based on the level of accuracy achieved using the default parameters.

CRAN Task View on ML & Stat Learning - 1

Several methods implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as ML. The functions can be roughly structured into the following topics:

- Neural Networks and Deep Learning: Single-hidden-layer neural network. Packages implementing deep learning flavours of neural
- Recursive Partitioning: Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book.
- Random Forests: The reference implementation of the random forest algorithm for regression and classification.
- Regularized and Shrinkage Methods: Regression models with some constraint on the parameter estimates.

CRAN Task View on ML & Stat Learning - 2

- Boosting and Gradient Descent: Various forms of gradient boosting.
- Support Vector Machines and Kernel Methods: flexible framework for kernel learning.
- Bayesian Methods: Bayesian Additive Regression Trees, Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian processes.
- Optimization using Genetic Algorithms: optimization routines based on genetic algorithms.
- Association Rules: provides both data structures for efficient handling of sparse binary data.

CRAN Task View on ML & Stat Learning - 3

- Fuzzy Rule-based Systems: implements a host of standard methods for learning fuzzy rule-based systems from data for regression and classification.
- Model selection and validation: function for hyper parameter tuning and function can be used for error rate estimation.
- Causal ML: ML framework in a variety of causal models.
- Visualization.
- XAI: Most packages and functions from the last section "Visualization" belong to the field of explainable artificial intelligence (XAI).
- Other procedures.

Generative AI (GAI) - 1

Overview of GAI

- GAI uses advanced ML techniques to analyze large datasets and generate new content based on the context, style, structure, and tone of the original data.
- When creating content, the AI model draws from patterns in the data to create outputs that are often indistinguishable from human-created material, whether it's text, images, code, or even music.

Generative AI (GAI) - 2

- As computational power and data availability increased, the introduction of more complex models marked significant milestones.
- Advancements in neural networks further improved the model's ability to understand language data, making interactions even more seamless and human-like.

Generative AI (GAI) - 3

Core concepts of GAI

- GAI is built upon the foundation of neural networks, which are computational models inspired by the structure and function of the human brain.
- These networks consist of multiple layers of interconnected neurons that process and transmit information, where each layer serves a specific purpose in processing information.

Generative AI (GAI) - 4

- The first layer, the input layer, receives raw data that gets transformed the more it travels from layer to layer, ultimately producing output in the final layer.
- This hierarchical structure allows neural networks to learn complex patterns and representations in data, with deeper layers often identifying more abstract patterns, just like how the human brain processes sensory information.
- Parameters, or weights, within these layers are crucial, as they determine how the input data is transformed.
  (»> nngraph)

Generative AI (GAI) - 5

How generative AI models work

- GAI models need to be trained on data. This involves exposing the neural network to large datasets, where the model learns to recognize patterns.
- The training process consists of two types of learning: supervised and unsupervised learning, and more.

Generative AI (GAI) - 6

- During training, the model learns to improve its predictions iteratively, refining its ability to generate coherent and relevant outputs.
- This is achieved through a process called back propagation, where the model makes predictions based on the input data, compares its outputs to the actual results, then calculates an error or loss. This feedback loop allows the network to fine-tune its parameters, gradually improving its ability to generate accurate and coherent content.

Generative AI (GAI) - 7

- In terms of language processing, generative AI breaks text down into smaller, manageable units called tokens, which can represent whole words, sub-words, or even individual characters, depending on the model's design.
- This allows the model to handle a diverse range of vocabulary and linguistic variations more effectively.
  Through token processing, the model can better understand relationships between words and generate more nuanced outputs, enhancing the model's ability to create coherent sentences and maintain context over longer passages, ultimately improving its performance in tasks like text generation and conversation.

Classical Stat Data Analysis vs. GAI era ML - 1

Differences between Classical Stat Data Analysis and GAI era ML

- Classical statistical data analysis emerged in an era of scarce and structured data.
- Methods were designed to extract meaning from small, well-organized data sets, under constraints of limited data availability and computational power.
- Statistical inference evolved to maximize insight from minimal input, relying heavily on assumptions of representativeness, structure, and stability.

- ML in GAI orients the assumptions and architecture of data analysis to accommodate the scale, diversity, representation, and expressiveness of modern data.

Classical Stat Data Analysis vs. GAI era ML - 2

1. **From relation-based to expression-based data**
   - Conventional data analysis rooted in structured, relation-based data, typically tabular data sets organized into variables and observations.
   - When working with unstructured data such as text, traditional approaches, such as text mining, typically began by extracting features and converting them into structured representations, before applying statistical or ML models.

Classical Stat Data Analysis vs. GAI era ML - 3

- Structured formats struggle to represent the richness of natural communication and perception, elements better preserved in expression-based data such as language, imagery, audio, and video.
  Expression-based data carries not only objective information, but also emotional tone, cultural context, ambiguity, and cognitive nuance.

- GAI engages this data directly.
- Models such as ChatGPT and DeepSeek process raw text, images, audio, and video in their native forms, learning representations that preserve contextual and semantic relationships without forcing data into predefined schemas.
- These systems interpret and generate content by synthesizing patterns across vast and heterogeneous sources.

Classical Stat Data Analysis vs. GAI era ML - 4

2. From sample-based inference to population-scale modeling

- Traditional statistics is built on sample-based inference, drawing conclusions about an unknown population from limited observations.
  This paradigm arose from necessity, as data were historically scarce and difficult to collect or process.
  Inferential tools such as hypothesis testing and confidence intervals were developed to quantify uncertainty when reasoning from small, imperfect samples.

- GAI reverses this logic. With access to vast training corpora, spanning books, articles, forums, and other large-scale repositories, models are trained on data sets intended to approximate broad populations of recorded human knowledge.

Classical Stat Data Analysis vs. GAI era ML - 5

- Rather than inferring the whole from a part ,
  GAI begins with population-scale representations and generates context-specific outputs in response to individual prompts.
  This reversal transforms not only the scale of analysis but also its underlying logic, reframing overfitting from a primary risk to a managed trade-off, and, in some cases, a necessary consequence of capturing complexity at scale.

Classical Stat Data Analysis vs. GAI era ML - 6

3. From parsimonious models to expansive representations

- Traditional statistical modeling has long favored parsimony and explicit uncertainty reasoning.
  Linear models, decision trees, and other interpretable techniques rely on compact functional forms, often paired with uncertainty quantification, to balance explanatory clarity with predictive reliability.
  The risk of overfitting is treated as a central concern and managed through variable selection, regularization, and carefully specified distributional assumptions.

Classical Stat Data Analysis vs. GAI era ML - 7

- GAI departs sharply from this paradigm.
  Large language models and other generative systems employ architectures with millions, or even billions, of parameters.
  Not to compressing data into a small set of key metrics, these models are designed to
  absorb, represent, and recombine information at scale.
  When training data are treated as a broad proxy for the world, overfitting is no longer framed as a primary failure mode, but as a trade-off, or even an inevitable consequence, of capturing complex, high-dimensional structure.

Potential resistance to adapt AI/ML - 1

Even though AI/ML expands the capabilities of the traditional stat data analysis, it also carries certain degrees of operational and philosophical resistances.

Potential resistance to adapt AI/ML - 2

Operational resistances

- ML algorithms is a multi-stage, multidisciplinary process.

  It needs a multidisciplinary team with the subject and ML domain expertise.
  Collaboration is necessary and valuable for both the data science and subject science components of the team,
- The choice and availability of data for algorithm for development and validation,
- Generalizability, reproducibility and replicability.

Potential resistance to adapt AI/ML - 3

Philosophical resistances

- Explainability:
  Interpretability - to describe the internals of a system in a way that is understandable to humans.
  Completeness - to describe the operation of a system in an accurate way.
- Trustworthiness: many aspects of the research tools that are not well or widely understood but continue to be used because they have been seem to work well, even if how or why remains opaque to many end users.

Potential resistance to adapt AI/ML - 4

European Union regulation considerations

- The use of algorithmic decisions in an increasingly wider range of applications has led researchers to caution against the rise of a "black box" society and demand increased transparency in algorithmic decision-making.
- Resistance to transparency includes the following and more:
  - Intentional organizational concealment on decision making procedures and without public scrutiny,
  - A mismatch between the mathematical optimization in high-dimensionality characteristic of ML and interpretation.
  - To ensure reproducibility and generalizability beyond the training data set.

Needs to explain outputs of ML - 1

- ML is at the core of many recent advances in science and technology.
  Whether humans are directly using ML classifiers as tools, or are deploying models within other products is important to consider.
  A vital concern remains:
  if users do not trust a model or prediction, they will not use it.
- It is important to differentiate between two different (but related) definitions of trust:
  (1) trusting a prediction, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and
  (2) trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed instead of seeing it as a black box.
  (»> go to SVM)

Needs to explain outputs of ML - 2

- Determining trust in individual predictions is an important problem when the model is used for decision making.
- Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it.
- To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest.
- Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics.
  In addition, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

Scopes to Model the Explanation Functions - 1

- Doshi-Velez and Kim [4] in their review paper defined: interpretability as the ability to explain or to present in understandable terms to a human.
- The need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation.

Scopes to Model the Explanation Functions - 2

The interpretability evaluation typically falls into two categories.

- First: evaluates interpretability in the context of an application: if the system is useful in either a practical application or a simplified version of it, then it must be somehow interpretable.
- Second: evaluates interpretability via a quantifiable proxy: a researcher might first claim that some model class are interpretable and then present algorithms to optimize within that class.
- Interpretability is used to confirm other important desiderata of ML systems, including
  the notions of fairness or unbiasedness, privacy, reliability,
  robustness, causality, usability, and trustworthy, etc.

Methods to model the Explanation Functions - 1

- Ribeiro, et al [9] define an explanation as a model $g \in G$, where G is a class of potentially interpretable models, such as linear models, or decision trees, i.e. a model $g \in G$ can be readily presented to the user with visual or textual artifacts.
- Let $\Omega(g)$ be a measure of complexity (as opposed to interpretability) of the explanation $g \in G$. For example, for decision trees $\Omega(g)$ may be the depth of the tree, while for linear models, $\Omega(g)$ may be the number of non-zero weights.
- Let the model being explained be denoted $f : R^d \rightarrow R$. In classification, $f(x)$ is the probability (or a binary indicator) that x belongs to a certain class.
- Further use $\pi_x(z)$ as a proximity measure between an instance z to x, so as to define locality around x.

Methods to model the Explanation Functions - 2

- Finally, let $L(f; g; \pi_x)$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x.
- In order to ensure both interpretability and local fidelity, we must minimize $L(f; g; \pi_x)$ while having $\Omega(g)$ be low enough to be interpretable by humans.
- The explanation produced by LIME is obtained by the following:

$$\eta_(x) = \text{argmin}_{g \in G} L(f; g; \pi_x) + \Omega(g). \quad (1)$$

This formulation can be used with different explanation families G, fidelity functions L, and complexity measures.

Classic Shapley Value Estimation - 1

- Shapley regression values are feature importance for linear models in the presence of multi-collinearity.
- Note on Shapley regression:
  - if data can be explained using linear models, one most likely does not need methods such as xgboost, randomForest, or deep learning, etc.
  - if data can't be well-explained by linear models, using Shapley regression is unlikely to completely explain the data.
  - most of real data involves variables having various degrees of inter-correlation, with complex co-linearity, and unknown interactions, the true model is difficult to discover.

Classic Shapley Value Estimation - 2

- The estimation method requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all features.
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature.
- To compute this effect, a model $f_{S\cup\{i\}}$ is trained with that feature present, and another model f_S is trained without that feature.
- Predictions from the two models are compared on the current input $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ - $f_S(x_S)$ where x_S represents the values of the input features in the set S.

Classic Shapley Value Estimation - 3

- Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S \subseteq F \setminus \{i\}$.
- The Shapley values are then computed and used as feature attributions. They are a weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (2)$$

Classic Shapley Value Estimation - 4

- For Shapley regression values, h_x maps 1 or 0 to the original input space, where 1 indicates the input is included in the model, and 0 indicates exclusion from the model.

- Shapley sampling values are meant to explain any model by: (i) applying sampling approximations to Equation (2), and (ii) approximating the effect of removing a variable from the model by integrating over samples from the training dataset.
- This eliminates the need to retrain the model and allows fewer than $2^{|F|}$ differences to be computed. Since the explanation model from Shapley sampling values is the same as that for Shapley regression values, it is also an additive feature attribution method.

Classic Shapley Value Estimation - 5

Lundberg & Lee [7] proved the following theorem:

Theorem: **only one possible explanation model** g **follows definition 1** and satisfies the Local accuracy, Missingness, and Consistency properties:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') \breve{} f_x(z' \setminus i)] \tag{3}$$

where $|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

Classic Shapley Value Estimation - 6

Then Lundberg & Lee [7] proposed

- SHAP values as a unified measure of feature importance.
- These are the Shapley values of a conditional expectation function of the original model; thus, they are the solution to Equation (3), where

$$f_x(z') = f(h_x(z')) = E[f(z)|z_S],$$

and S is the set of non-zero indexes in z'.

A clinical example - Stomach (gastric) cancer

Medical images of gastric cancer

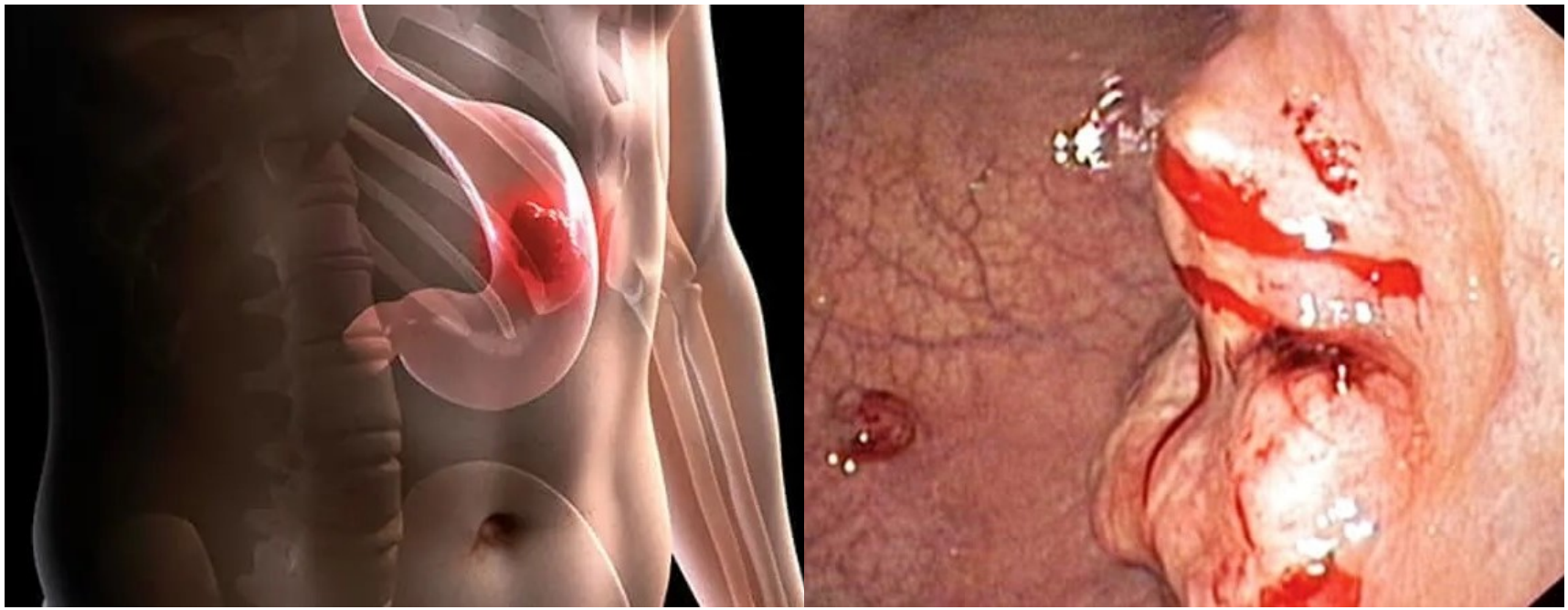


Figure 5: Stomach (gastric) cancer

Description of clinical study - 1

A Clinical Protocol: A randomized, multicenter, open-label, phase 3 study of EXP+chemo vs chemo in subjects with previously untreated advanced or metastatic Stomach Cancer (SC)

Objectives:

- Primary Objectives - EXP+chemo vs chemo: to compare OS and PFS via a 2-arm trial

Results of data analysis - 1

The xgboost computations

```
** Variables selected from databases - (adsl, efficacy, lab)

** Baseline values of input variables to run XGB
(only sample 100 cases as illustration)

 $ X        :'data.frame':        100 obs. of  24 variables:
  ..$ ECOGPSN : num [1:100] 0 0 0 1 1 0 1 1 0 1 ...
  ..$ LIVERMEN: num [1:100] 0 0 0 0 0 0 0 1 1 1 ...
  ..$ MSIN    : num [1:100] 2 2 2 2 2 2 2 2 2 2 ...
  ..$ TRT01AN : num [1:100] 3 2 2 1 1 3 3 2 3 2 ...
  ..$ CPDL1N  : num [1:100] 2 1 2 2 2 2 2 2 2 2 ...
  ..$ HBCATN  : num [1:100] 2 2 2 2 2 2 2 1 2 2 ...
  ..$ LAUCLASN: num [1:100] 2 1 2 2 1 2 2 1 2 1 ...
  ..$ ALB     : num [1:100] 35 42 38 47 37 42 38 39 46 47 ...
  ..$ ALP     : num [1:100] 155 100 192 130 121 58 89 139 199 130 ...
  ..$ AST     : num [1:100] 15 14 17 17 34 20 26 20 21 50 ...
  ..$ CA      : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
  ..$ CREAT   : num [1:100] 56 94 50 93 81 55 122 106 59 66 ...
```

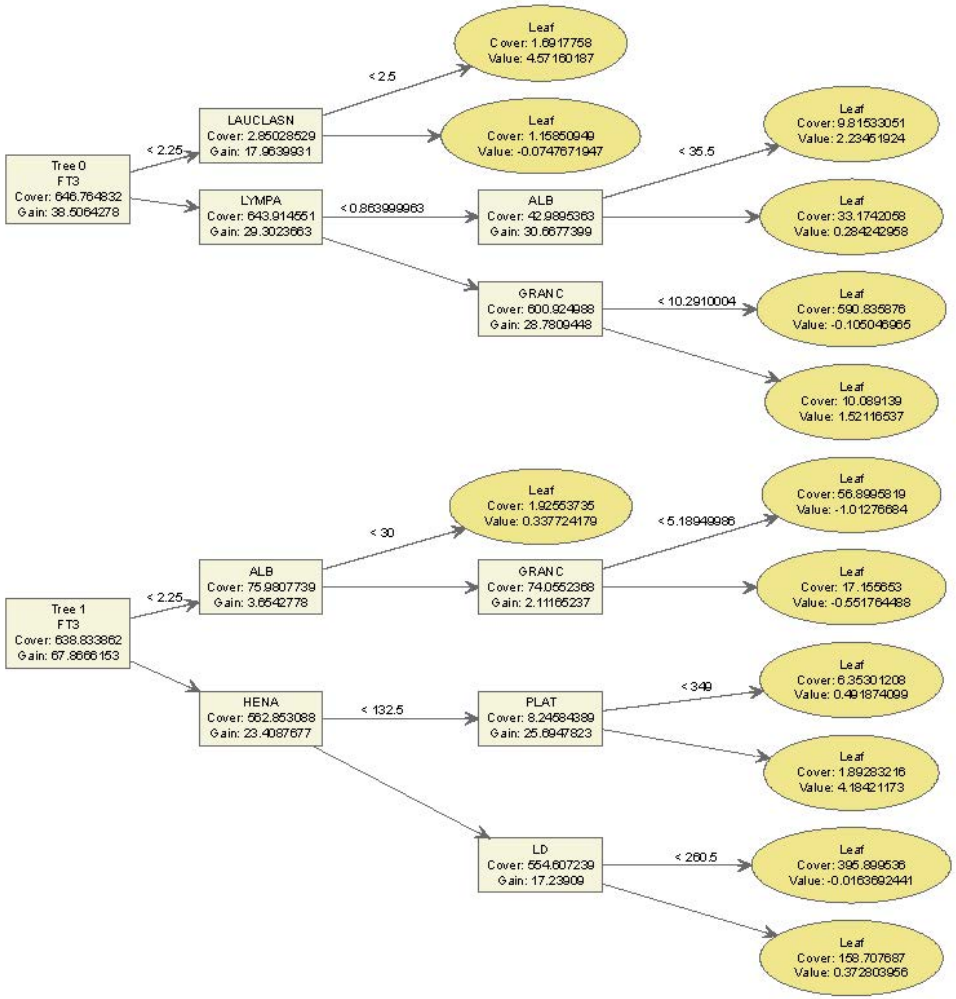
Results of data analysis - 2

```
..$ FT3      : num [1:100] 3.7 3.6 3.6 3 4.4 3.8 2.9 5.6 4.3 5.1 ...
..$ FT4      : num [1:100] 12.9 16.4 14.1 16 16.4 12.9 16.5 16.5 18.1 19.2 ...
..$ GRANC    : num [1:100] 3.73 6.35 2.04 3.89 5.07 ...
..$ HB       : num [1:100] 125 112 123 154 135 127 147 75 138 148 ...
..$ HECA     : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
..$ HENA     : num [1:100] 143 138 141 142 135 141 132 137 140 140 ...
..$ HOCA     : num [1:100] 2.28 2.35 2.28 2.44 2.18 2.3 2.4 2.3 2.33 2.47 ...
..$ HONA     : num [1:100] 143 138 141 142 135 141 132 137 140 140 ...
..$ LD       : num [1:100] 149 127 185 279 99 150 182 196 154 455 ...
..$ LYMPA    : num [1:100] 1.921 1.14 0.787 1.7 3.56 ...
..$ PLAT     : num [1:100] 320 275 188 232 346 370 477 634 174 172 ...
..$ TBILI    : num [1:100] 5.1 6 13.7 18.5 6.5 7 6.5 5.6 17.1 10.5 ...
```

Results of data analysis - 3

```
call:
xgb.train(params = param, data = dtrain, nrounds = 2, watchlist = watchlist)
params (as set within xgb.train):
max_depth = "3", eta = "1", verbose = "0", nthread = "2",
objective = "survival:cox", eval_metric = "cox-nloglik",
validate_parameters = "TRUE"
xgb.attributes: niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 24
niter: 2
nfeatures : 24
evaluation_log:
 iter train_cox_nloglik eval_cox_nloglik
    1          5.980229         5.980229
    2          5.973864         5.97386
```

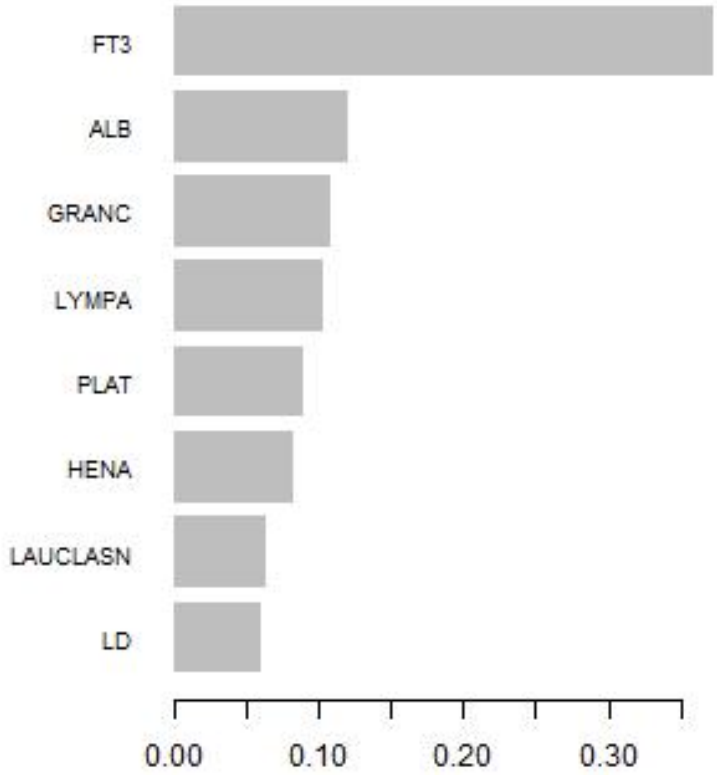
Results of data analysis - 4



Results of data analysis - 5

	Feature	Gain	Cover	Frequency
1:	FT3	0.37298139	0.3337465994	0.18181818
2:	ALB	0.12034509	0.0308851722	0.18181818
3:	GRANC	0.10832033	0.1752275935	0.18181818
4:	LYMPA	0.10274442	0.1671628190	0.09090909
5:	PLAT	0.09009496	0.0021406544	0.09090909
6:	HENA	0.08207939	0.1461189359	0.09090909
7:	LAUCLASN	0.06298809	0.0007399456	0.09090909
8:	LD	0.06044632	0.1439782800	0.09090909

Results of data analysis - 6



SHAP analysis - 1

The SHAP computations following xgb outputs

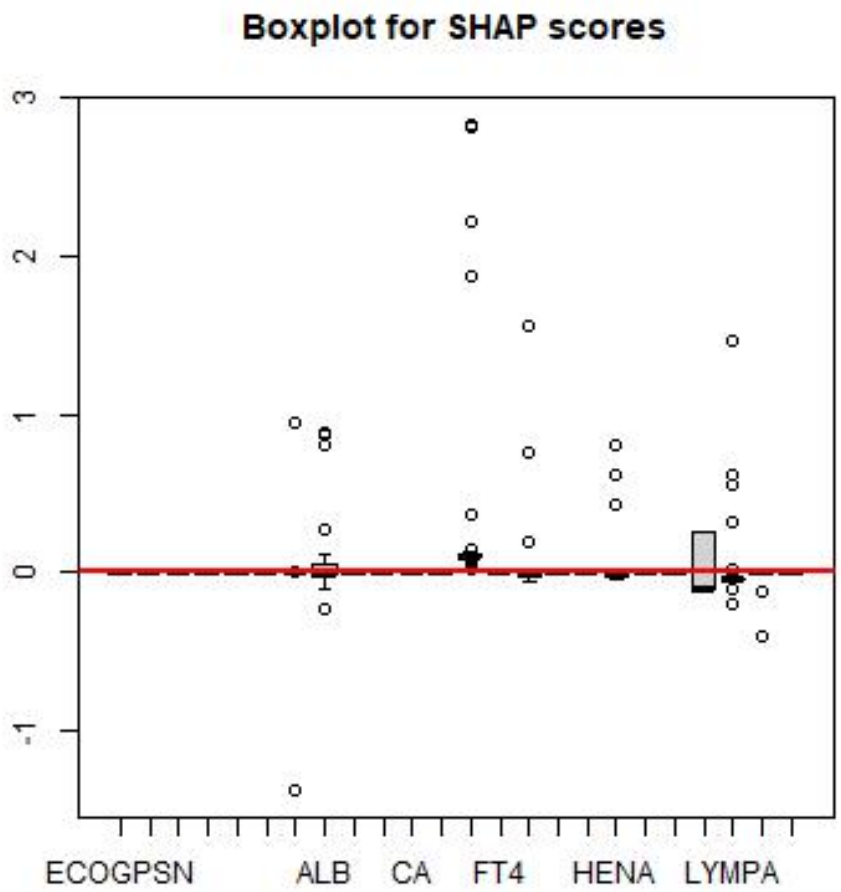
```
** SHAP importance scores of input variables for each case

'data.frame':   100 obs. of  24 variables:
 $ ECOGPSN : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LIVERMEN: num  0 0 0 0 0 0 0 0 0 0 ...
 $ MSIN    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TRT01AN : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CPDL1N  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ HBCATN  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LAUCLASN: num  0.00416 0.00416 0.00416 0.00416 0.00416 ...
 $ ALB     : num  0.0482 -0.0165 -0.2391 -0.0167 -0.0167 ...
 $ ALP     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AST     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CA      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CREAT   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ FT3     : num  0.1015 0.0742 0.1026 0.1245 0.1014 ...
 $ FT4     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ GRANC   : num  -0.03258 -0.00563 -0.01895 -0.03258 -0.03258 ...
```

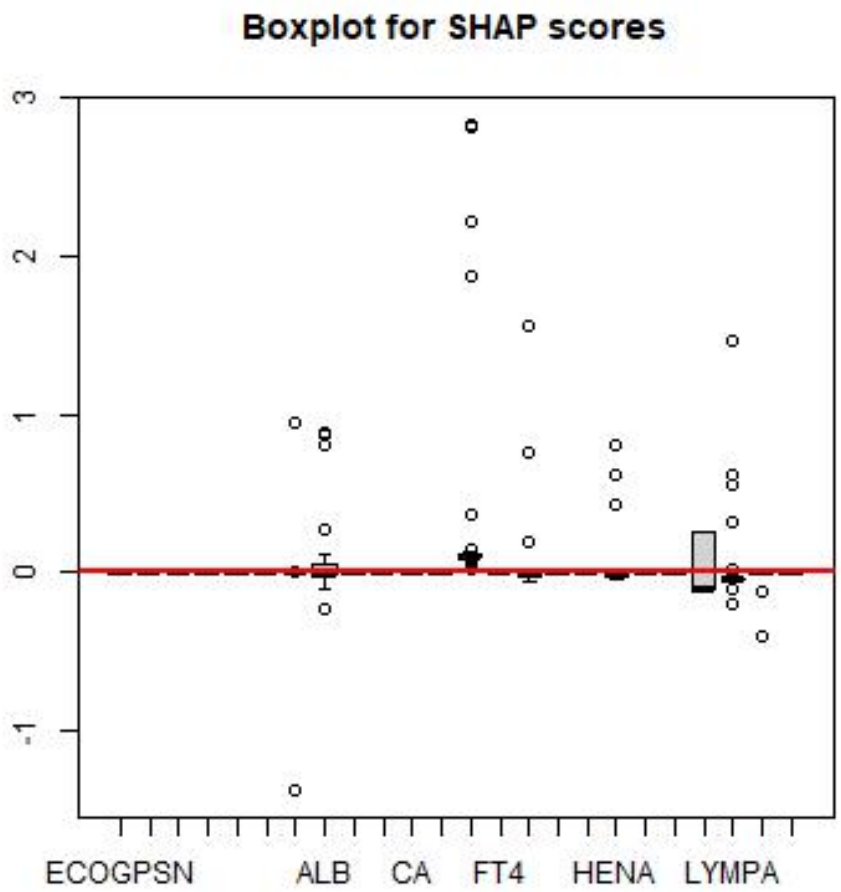
SHAP analysis - 2

```
$ HB      : num  0 0 0 0 0 0 0 0 0 0 ...
$ HECA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ HENA    : num  -0.01197 -0.01197 -0.01197 -0.00923 -0.01197 ...
$ HOCA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ HONA    : num  0 0 0 0 0 0 0 0 0 0 ...
$ LD      : num  -0.104 -0.104 -0.104 0.259 -0.104 ...
$ LYMPA   : num  -0.1049 -0.0398 0.557 -0.0398 -0.0398 ...
$ PLAT    : num  -0.00572 -0.00572 -0.00572 -0.00572 -0.00572 ...
$ TBILI   : num  0 0 0 0 0 0 0 0 0 0 ...
```

SHAP scores via xgb model



SHAP scores via xgb model



Distributions of SHAP importance of FT3 and LYMPH

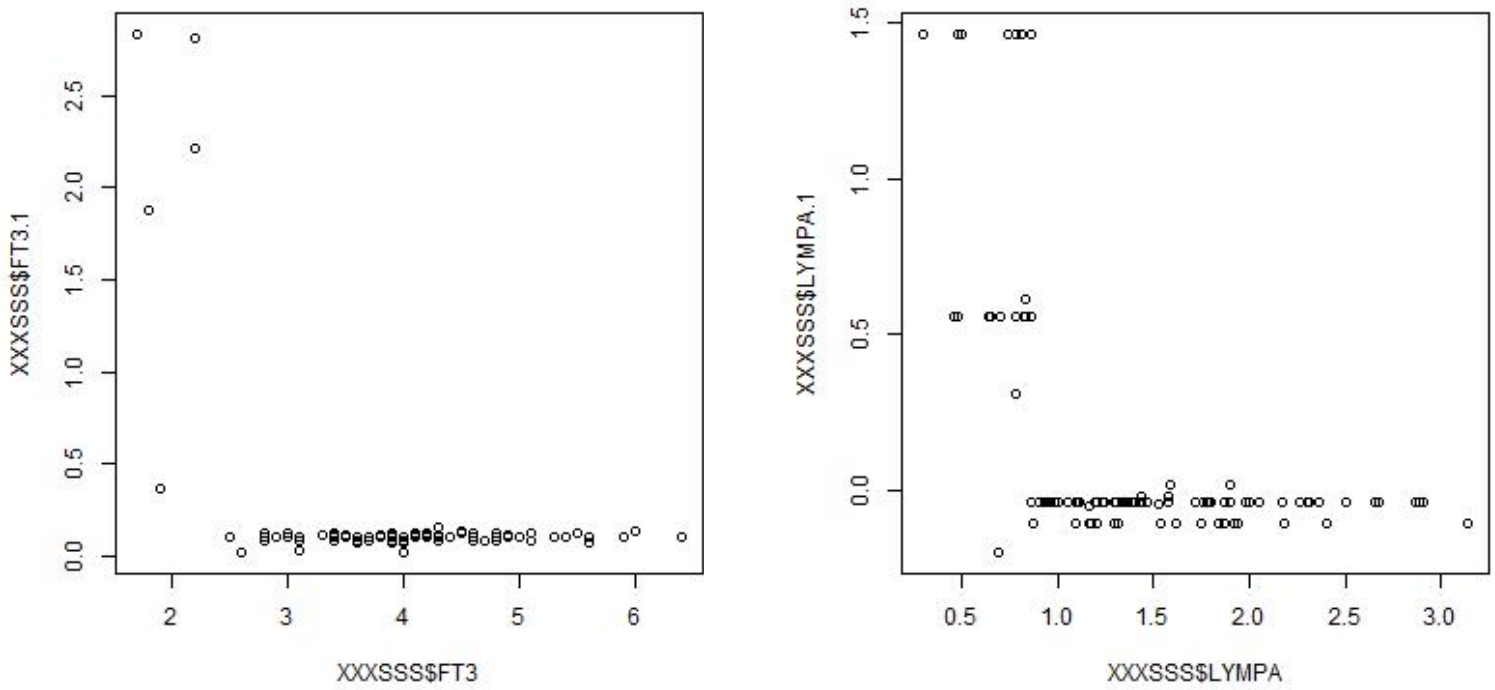
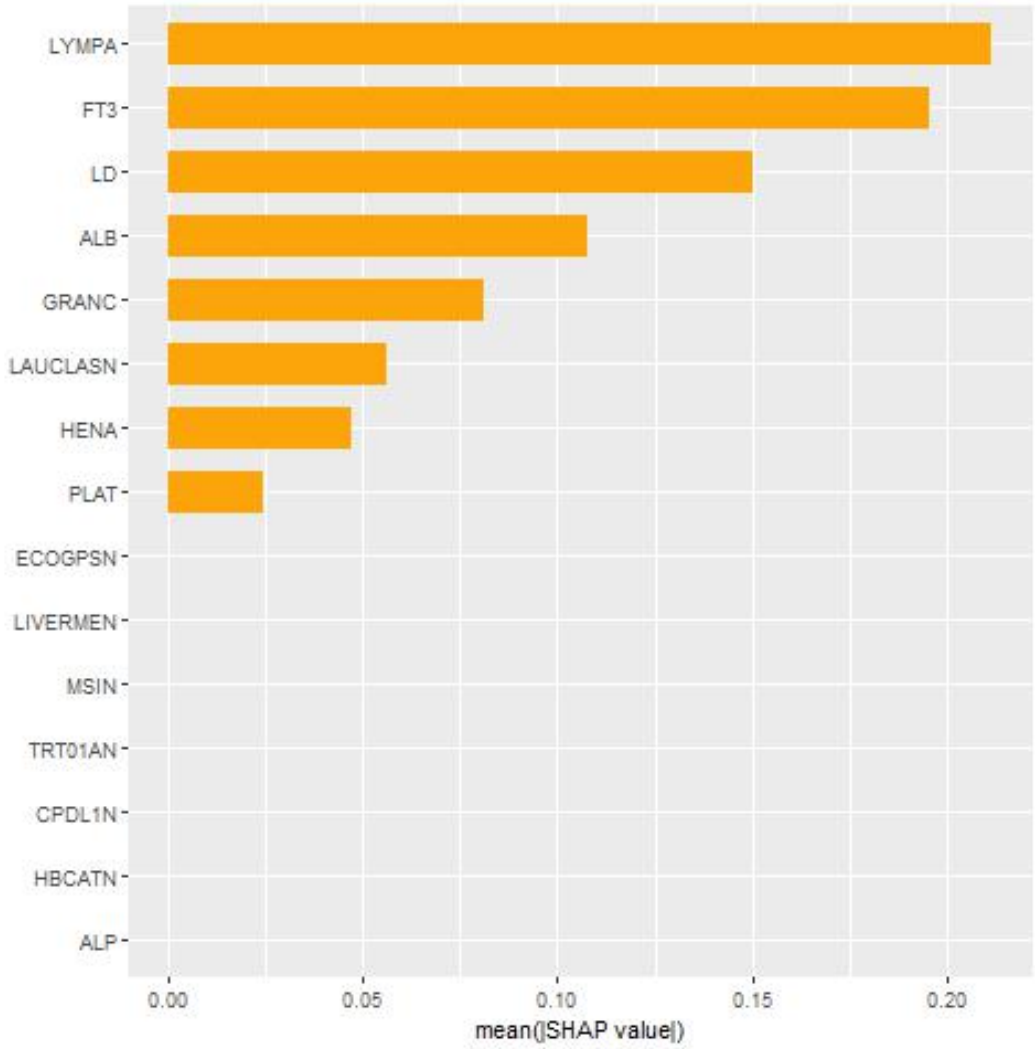
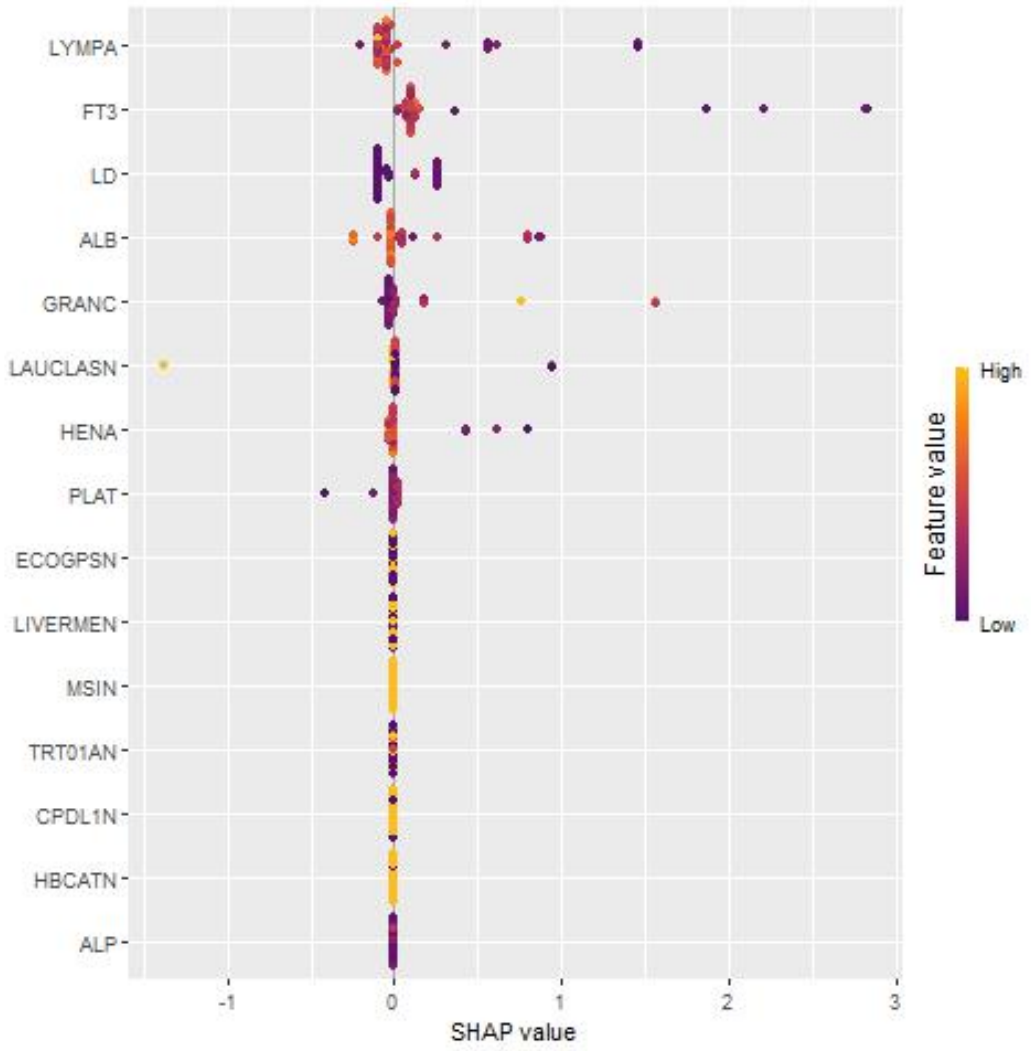


Figure 6: Distributions of SHAP importance of FT3 and LYMPH

SHAP (ordered by means) via xgb model



SHAP scores via xgb model



Applications of AI/ML on clinical research - 1

Planning and conduct of clinical trials

- Define trial objectives
- Trial designs
- Study protocol optimization.
- Data collection and management: experimental designs, commercial/government databases, wearable and mobile devices, social media, Internets, etc.
- Trial execution: patient recruitment, patient retention, trial compliance, trial risk management, etc.

Applications of AI/ML on clinical research - 2

§ **Patient populations selection and patient recruitment**

- Selection of specific patient populations for test compound.
- Increased homogeneity of trial patients may decrease the sample size required for significant treatment effect.
- Following example of patient selection for proper trial from Cincinnati Children's Hospital.

Applications of AI/ML on clinical research - 3

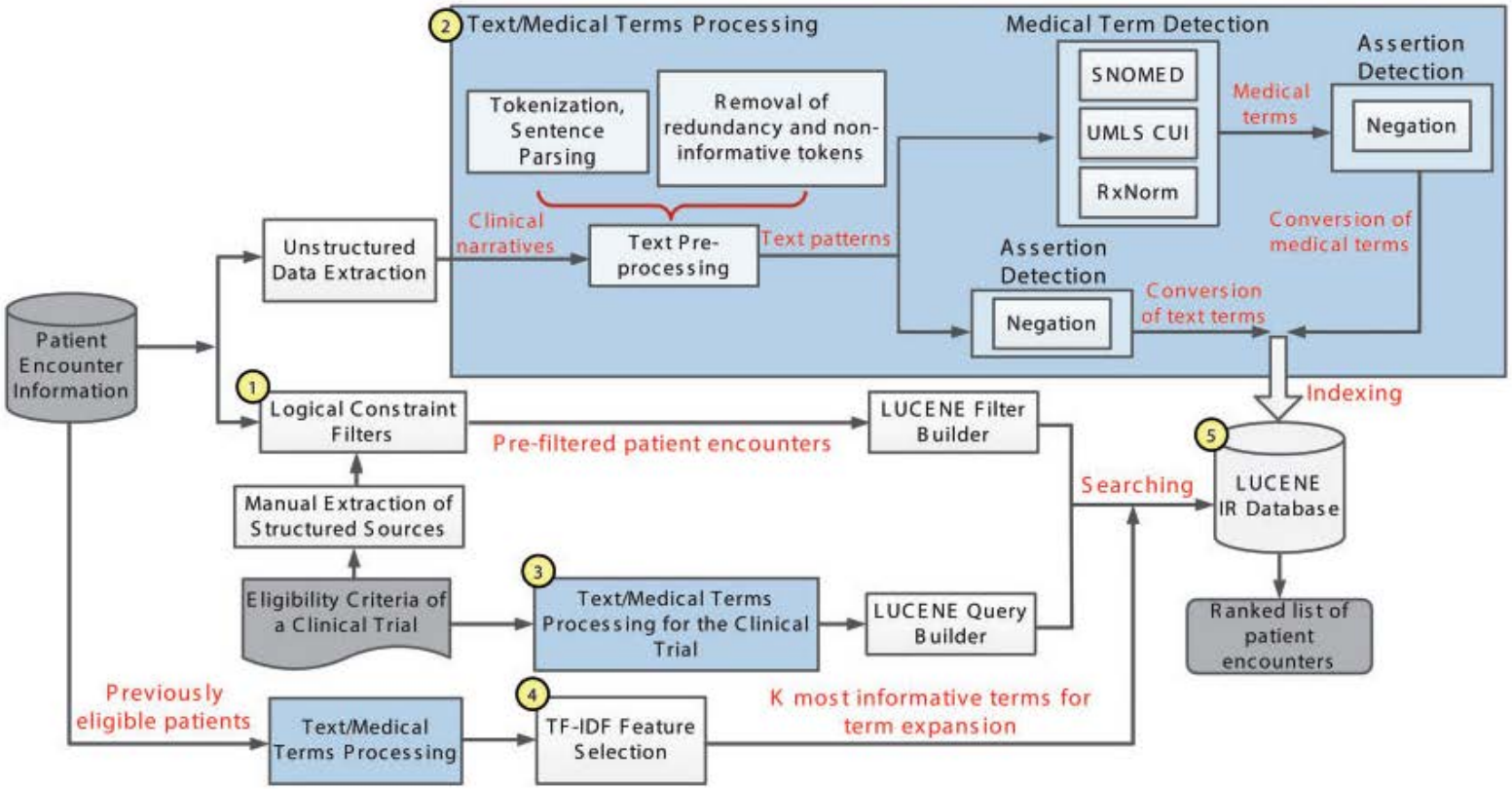


Figure 7: Automated screening for trial recruitment

Applications of AI/ML on clinical research - 4

§ **Data analysis**

- data may come from several sources, such as
  - designed experiment: such as controlled clinical trials, designed sample survey, etc., usually with more structured specifications.
  - observational data
  - data from wearable and mobile devices, etc.
- Analytical methodologies
  - fundamental statistical knowledge in estimation, hypothesis testing, prediction, and modeling, etc., is essential.
  - classical statistical methods and ML are both needed in EDA, estimation, testing, prediction, and modeling, etc.

Applications of AI/ML on clinical research - 5

- Efficacy data analysis
  - the commonly used methods including Cox proportional hazard regression, logistic regression, and multiple regression, etc.
  - ML methods for data from inter-current events and observational studies, which usually do not preserve the randomization property, are very useful.

Applications of AI/ML on clinical research - 6

- Safety data analysis
  - understanding of the safety aspect of the test compounds is critical, however, it is usually under supported by statisticians.
  - to understand the relationship between various safety events, or between safety events and efficacy, the conventional tabulations are insufficient;
  - more insightful methodologies are available in ML toolbox, for example, the network analysis as shown below.

Applications of AI/ML on clinical research - 7

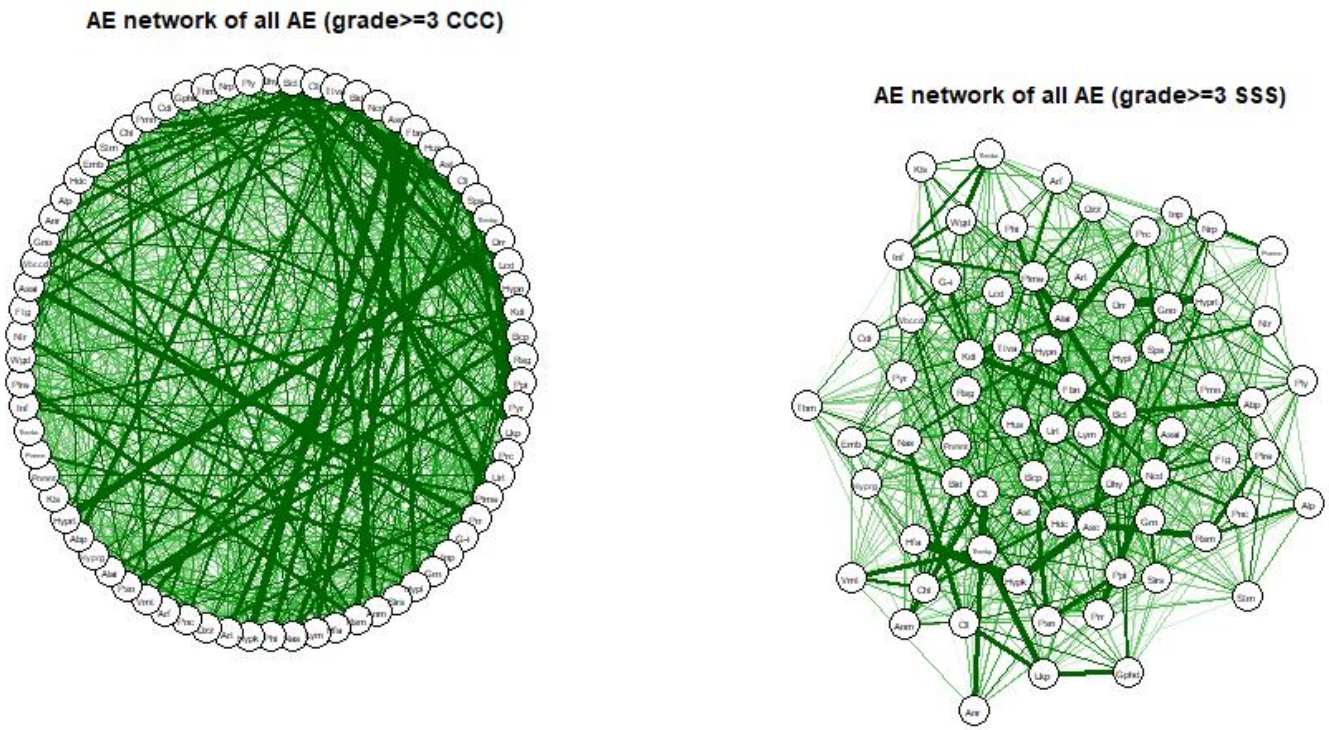


Figure 8: Network analysis of serious adverse events

Applications of AI/ML on medical devices - 1

FDA guidelines of Good ML Practices on medical devices

- Multi-disciplinary expertise is leveraged throughout the total product life cycle
- Good software engineering and security practices are implemented
- Clinical study participants and data sets are representative of the intended patient population
- Training data sets are independent of test sets
- Selected reference datasets are based upon best available methods
- Model design is tailored to the available data and reflects the intended use of the device
- Focus is placed on the performance of the human-AI team
- Testing demonstrates device performance during clinically relevant conditions

Applications of AI/ML on medical devices - 2

- Users are provided clear, essential information
- Deployed models are monitored for performance and re-training risks are managed

Summary: the expanded path ahead of us 1

- During the past few decades since Professor Peace launched the BASS, multiple scientific areas has great advances, which includes clinical research, AI and ML, and statistical computing.
- Traditionally, our data for clinical recherche are mostly numerical, however, data nowadays includes text, image, audio, etc., especially when we deal with real-world data.
- As the most important data analytical group in this industry, we need to extend our scopes and skill sets to accommodate these changes and enhance our importance in the field of clinical and health research.

- Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) Classification and Regression Trees. Wadsworth.
- Chen, T and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System, 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.
- CRAN Task View: ML & Statistical Learning.
- Doshi-Velez, F. and Kim, B () Towards A Rigorous Science of Interpretable ML
- B. Goodman, B. and Flaxman, S. (2016) "European union regulations on algorithmic decision-making and a "right to explanation"," arXiv:1606.08813.
- Huber, J. P. (1997) Speculations on the path of Statistics. In: *The Practice of Data Analysis. Essays in Honor of John W. Tukey.* Brillinger, D.R., Fernholz, L. T. and Morgenthaler, S. (eds). Princeton University Press, NJ.

- Lundberg, S. & Lee, S (2017) A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- Mitchell, T. (1997). Machine Learning. McGraw Hill.
- Ribeiro,M., Singh,S. , Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016 San Francisco, CA, USA
- Tsai, K.T. Assessing Regression Modeling with Ordinal Responses. *Presentation at the Joint Statistical Meetings of the American Statistical Association* 2008.
- Turing, Alan (October 1950). "Computing Machinery and Intelligence." *Mind*,v59, 236.

Analytic of AI/ML



Figure 9: Analytic of AI/ML (»> retAnalytic)

Neural Network example

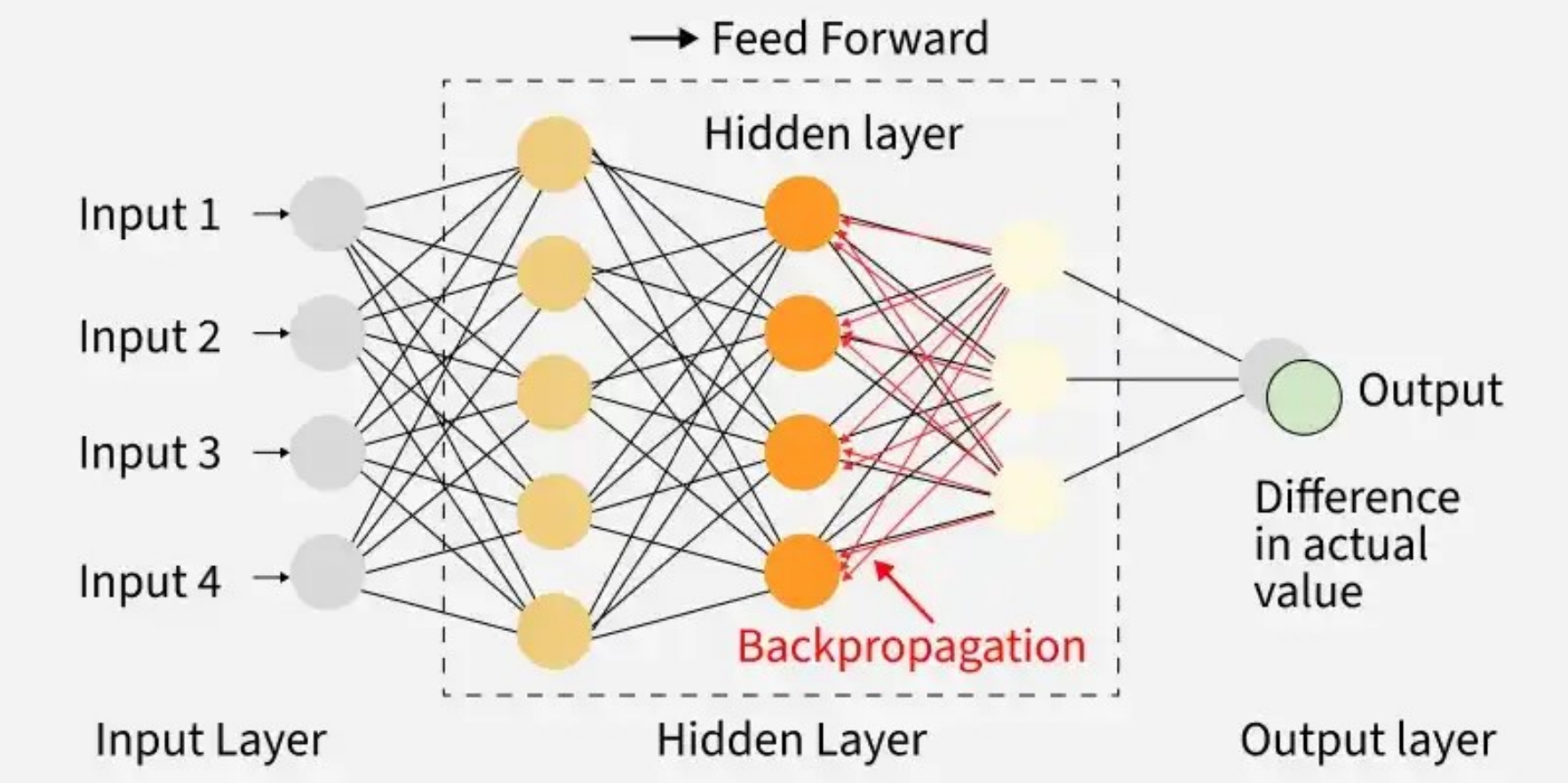


Figure 10: Neural Network example (»> returnNN)

Classification example

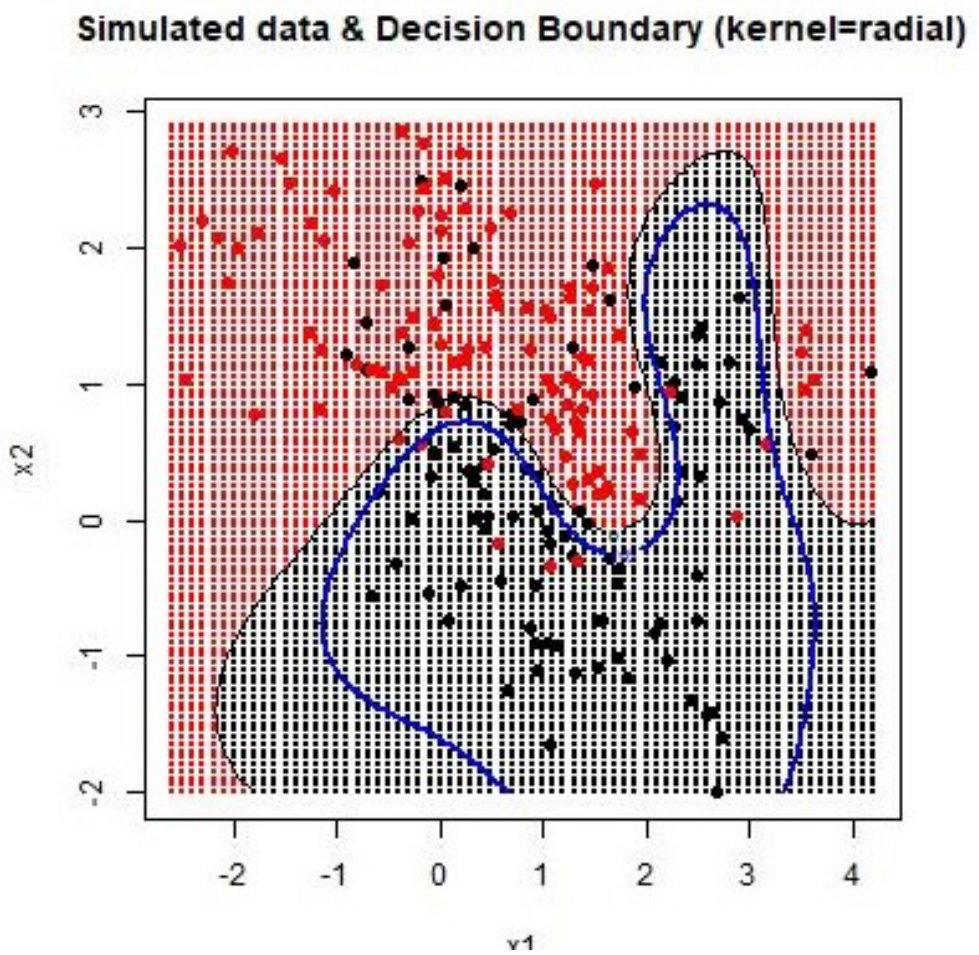


Figure 11: Classification example (»> back2SVM)