Model complexity vs interpretability

Can you have it all?

Lucia Pagani, PhD – Senior Data Scientist

Today's journey

- High level overview of different techniques
  - Algorithms for model interpretation (XAI)
  - Wrappers for framework
- Hands-on example of few most common methods for interpreting models

IQVIA CORE

Domain Expertise

+1,100 Medical Doctors

+1,400 PhDs

+2,500 Statisticians

+850 Epidemiologists / RWI experts

Transformative Technology

+100,000 users on our software platforms

E360 salesforce

Unparalleled Data

+530m non-identified patient records

+800,000 data sources

+75b healthcare records

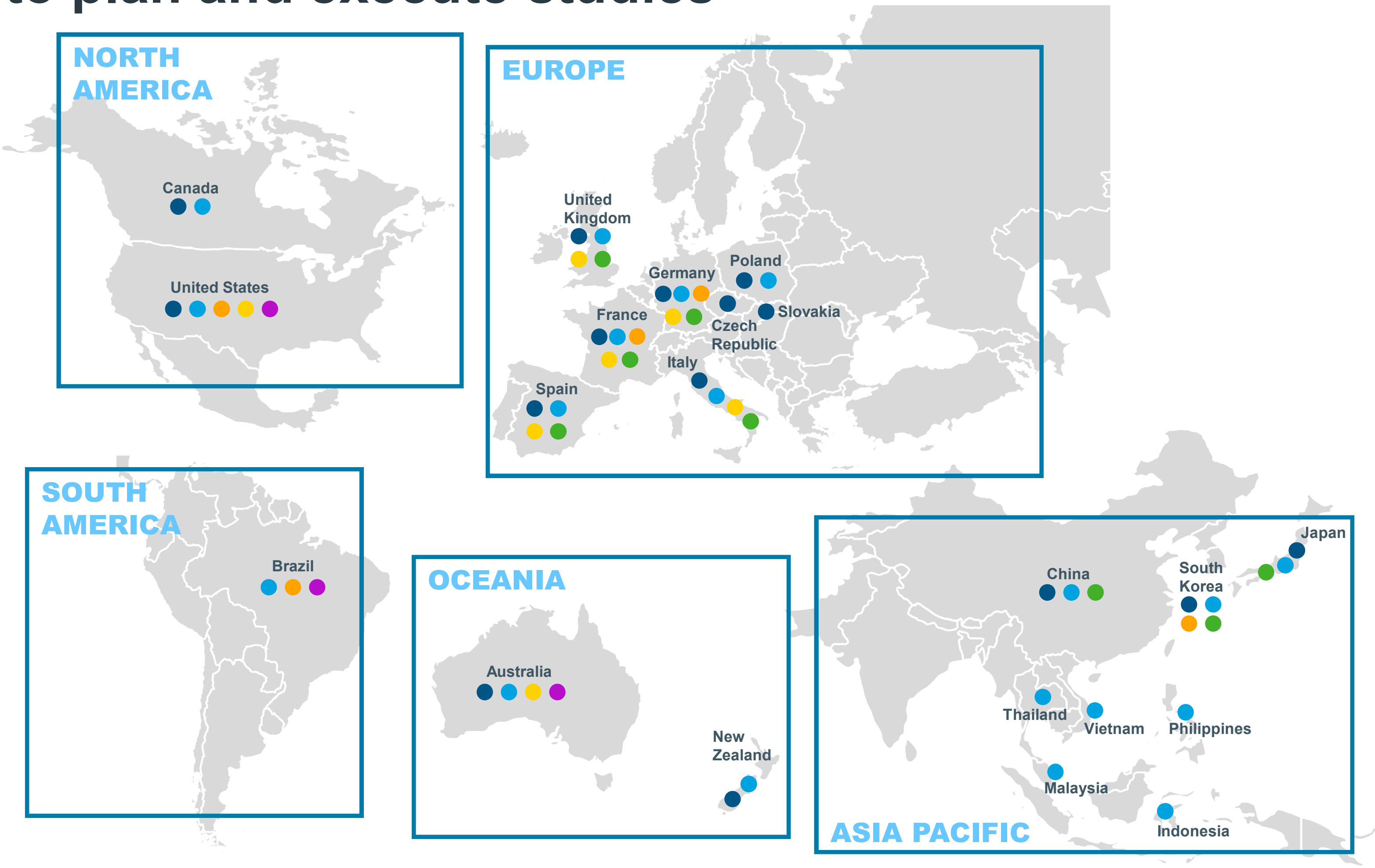
+85% of global drug sales covered

Advanced Analytics

+200 Patents and patents pending

+6,000 client engagements / yr

IQVIA uses insights from real-world data assets around the globe to plan and execute studies



IQVIA Data Assets

- ● Hospital or Non-Retail Pharmacy
- ● Retail Pharmacy
- ● Medical Claims
- ● Electronic Medical Records (EMR)
- ● TA-Specific*
- ● Other (Lab, OTC or Ethnicity)

*Oncology cross-sectional survey- CN, DE, ES, FR, IT, JP, KR, UK
Oncology EMR-U.S.

What is the Analytics Center of Excellence (ACOE)?

Site ID

48% of trial sites miss enrollment targets

Design

60% of trials have a protocol amendment

Recruitment

80% of trials delayed, mainly due to enrollment

Analytics Center of Excellence

A **global** unit, with **local country presence** to fully leverage the data, methods, and expertise specific to each country

A team dedicated to optimize clinical research, from clinical development through real-world evidence (RWE)

An enabler for the Data driven study execution

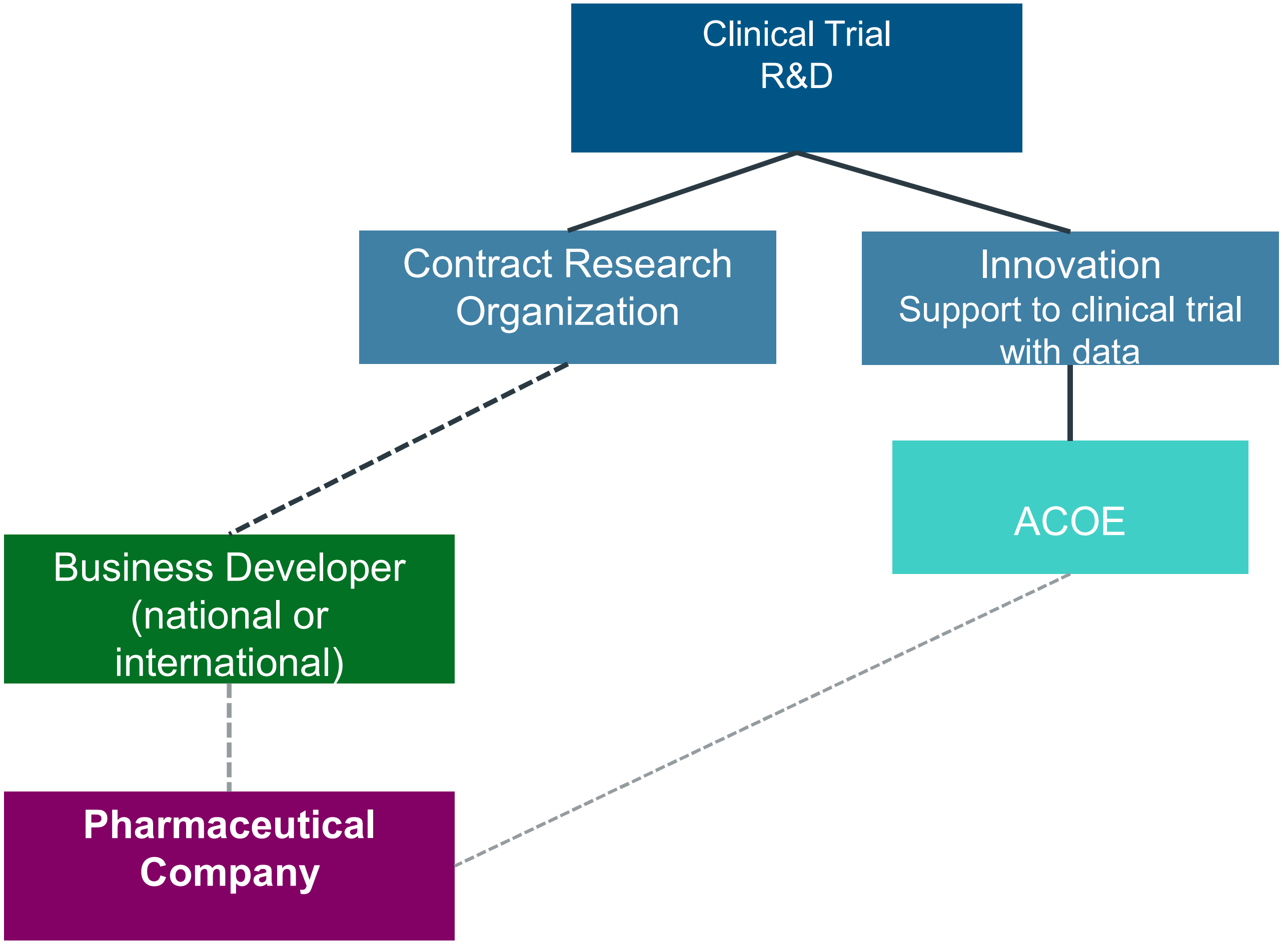
A differentiated capability

Data-Driven Patient Recruitment Services

Integrated to enhance the broader recruitment strategy

SERVICE	ACTIVITY	OUTCOME
Enrollment Predictability **Evidence-based recruitment plans** informed by site-level patient density and high potential referral locations	CRA-led data-driven recruitment goal and strategy review	**Site-level recruitment action planning** for more accurate enrollment goal planning
Enrollment Trends **Action plan for low/no enrollers** by comparing site-level screening activity with recent eligible patient activity	Frequent, CRA-led recruitment action planning with refreshed data	**Minimize low/non enrolling sites** **Intervene earlier** to maximize site-based and supplemental recruitment
Referral Network Service **Referral strategy to deliver more patients per site** leveraging insights on nearby providers with eligible patients, referral patterns and institutional relations	SPN or CTE-led high-touch referral outreach support	**Proactively increase sites' access to eligible patients** by mobilizing referral connections with community-based physicians
Patient-centric Engagement **Consumer and social insights** that inform patient profiles, messaging approach and channel, optimizing direct-to-patient recruitment media spend and impact	Online behavioural and attitudinal analysis; demographic data	**Precise online targeting** to locate interested patients with cost-effective outreach strategies

Stakeholders



Clinical Trial
R&D

Contract Research Organization

Innovation
Support to clinical trial with data

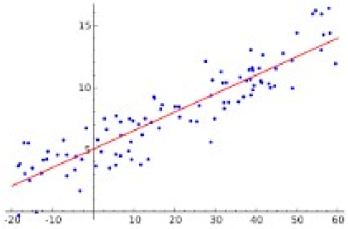
ACOE

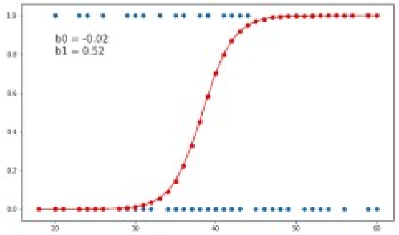
Business Developer (national or international)

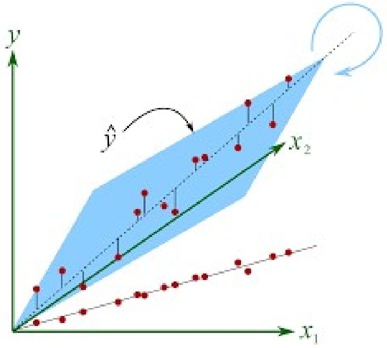
Pharmaceutical Company

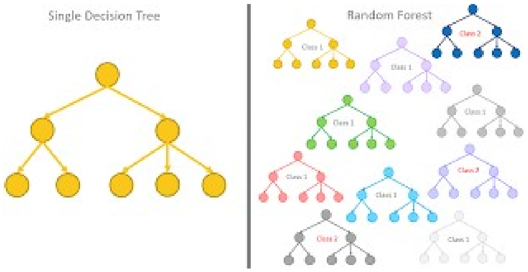
Our models span throughout the whole range of complexity

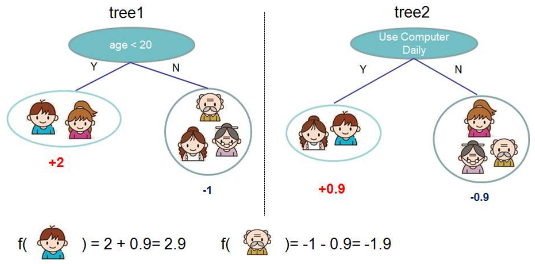
As simple as Linear Regression and as complex as multi-input Deep Neural Network

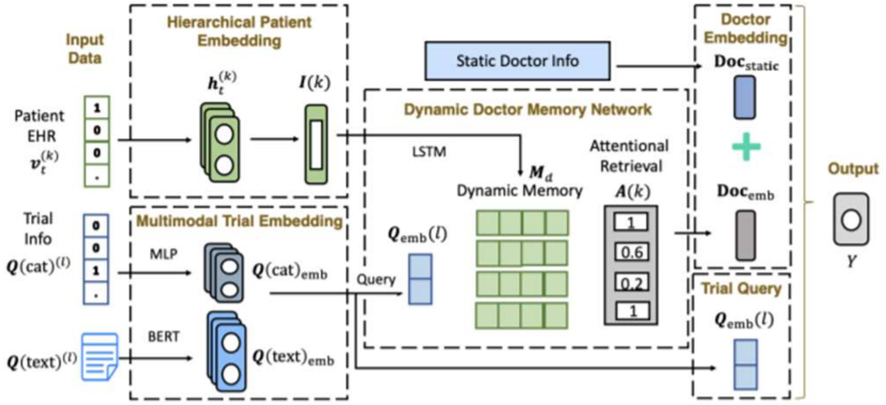






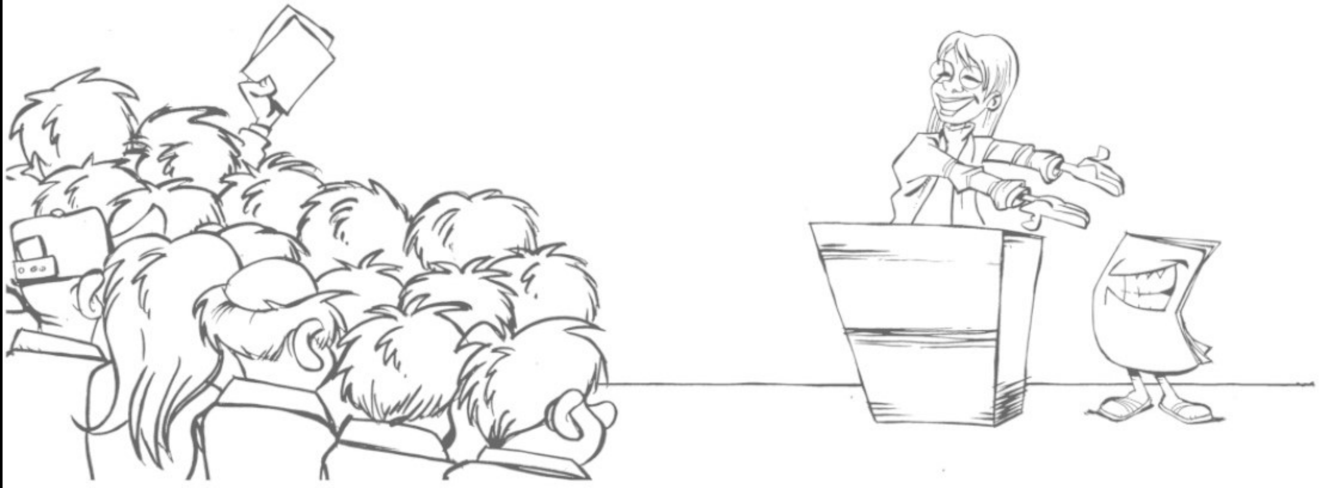






Performance versus interpretability

Difference between academia and industry



What clinicians want

Contextualizing explainable machine learning for clinical end use

- Explainability as a mean to justify their clinical decision making
- Understand clinically relevant model features that align with current evidence-based medical practice
- Explainability facilitates trust in the model, directing choice of patient population
- Clinicians expect to see patient specific variable importance and population level variable importance

Tonekaboni et al, 2019, https://arxiv.org/pdf/1905.05134.pdf

Do you trust AI?

Would you trust AI with 98% performance?

… or you would rather trust human with 85% performance?

Questions answered by interpretation

The whats, the whys, and the hows

- **What drives model predictions?**
  - Which features are important in the decision-making policies of the model
  - Translates with **fairness** of the model

- **Why did the model take a certain decision?**
  - Validate and justify why certain key features were responsible in driving certain decisions
  - Ensures accountability and **reliability** of the model

- **How can we trust model predictions?**
  - Evaluate and validate any data point and how a model takes decisions on it
  - Translates in **transparency** of the model

Interpretability

- For the **developer**
  - Improve performance
  - Model debugging
    - troubleshoot data leakage, bias, or feature engineering
  - Model validation
- For the **user**
  - judicial sentencing risk scores
  - credit scoring or fraud detection
  - health assessment
  - loan lending
- For the **stakeholders**
  - Reasoning behind a model's result to automate a manual process
    - Clinical study site identification
  - **Knowledge discovery**
    - Understanding why a group of subjects were predicted to develop a disease may unravel hidden pattern on new prodromic factors of the disease

Types of interpretation

Interpretation

Intrinsic (white box)
or
Post-hoc (apply interpretability methods after the training)

Model specific (eg coefficients or rules)
or
Model agnostic (eg input perturbation)

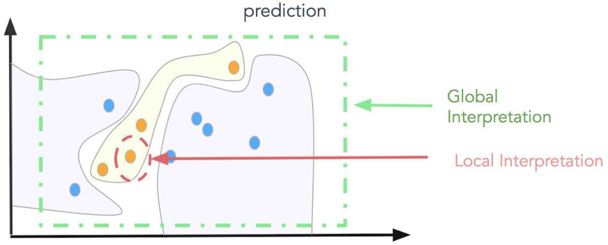
Global interpretation
or
Local interpretation

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



datascience.com

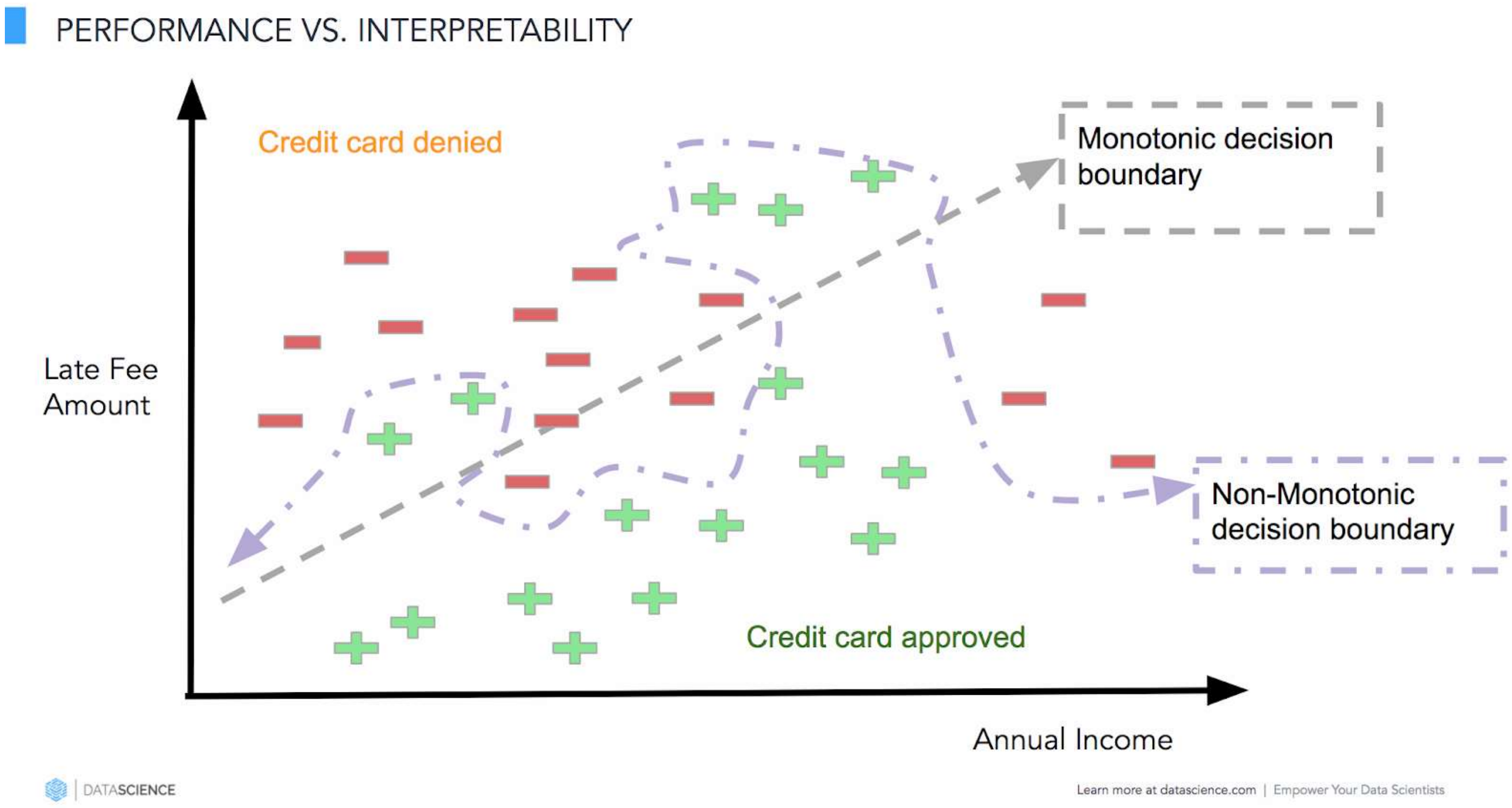
Interpretation techniques for Machine Learning algorithms

White box models

Intrinsic interpretability

- Linear models
  - Ordinary Least Squares – coefficients
  - Ridge – coefficients are penalized (l1-norm), so less important tend to 0
  - Lasso – coefficients are penalized (l2-norm) so it estimates sparse coefficients so less important are 0
  - Elastic net – coefficients are l1-norm and l2-norm penalized: the best of the two worlds
  - Logistic Regression – can penalize coefficients
- Linear kernel SVM
- Decision trees

Why do we need complex models?



PERFORMANCE VS. INTERPRETABILITY

Monotonic decision boundary

Non-Monotonic decision boundary

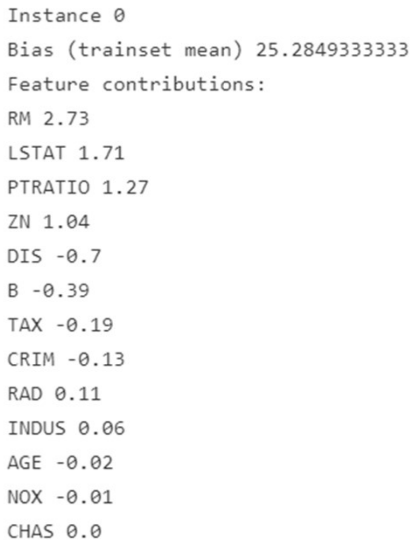
DATASCIENCE

Learn more at datascience.com | Empower Your Data Scientists

Average decision tree predictions

TreeInterpreter

- Decomposing each prediction into bias and feature contribution components
- It works for sklearn decision trees
- Combining intermediate values in all nodes in decision trees, it is possible to extract the prediction paths for each individual prediction and decompose the predictions via inspecting the paths to give leaf node_ids for predictions
- Clearly understanding why two predictions are different
- It can work as a local and global interpretation technique

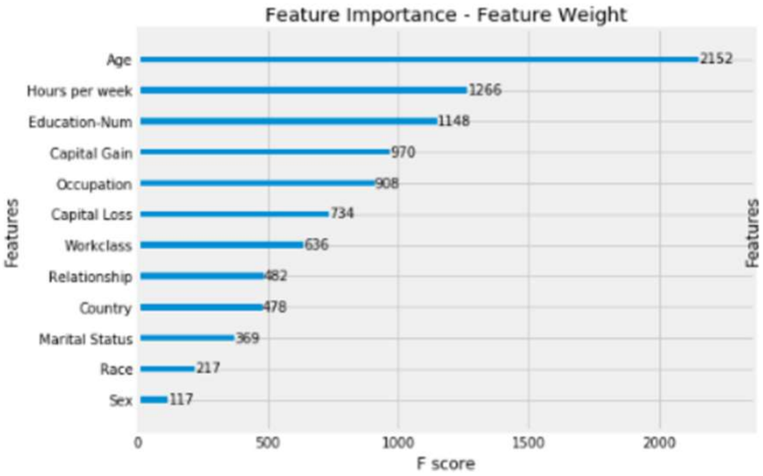


```
Instance 0
Bias (trainset mean) 25.2849333333
Feature contributions:
RM 2.73
LSTAT 1.71
PTRATIO 1.27
ZN 1.04
DIS -0.7
B -0.39
TAX -0.19
CRIM -0.13
RAD 0.11
INDUS 0.06
AGE -0.02
NOX -0.01
CHAS 0.0
```

https://github.com/andosa/treeinterpreter

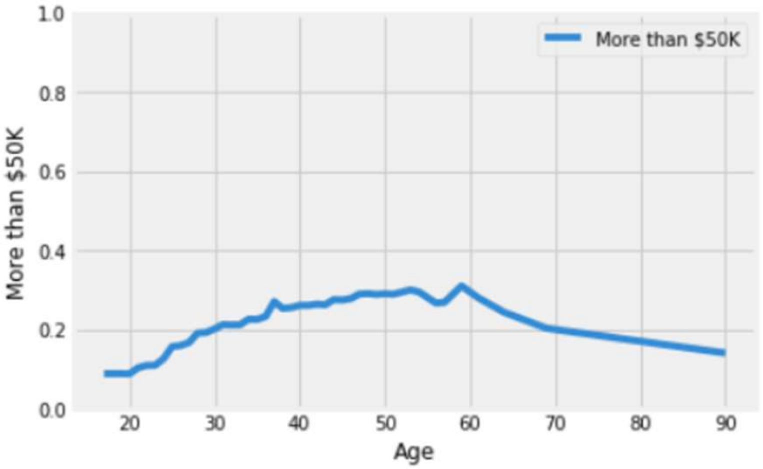
Feature importance

- Global interpretation method
- Feature importance is generic term for the degree to which a predictive model relies on a particular feature.
- **Feature Weights:** This is based on the number of times a feature appears in a tree across the ensemble of trees
- **Gain:** This is based on the average gain of splits which use the feature
- **Coverage:** This is based on the average coverage (number of samples affected) of splits which use the feature



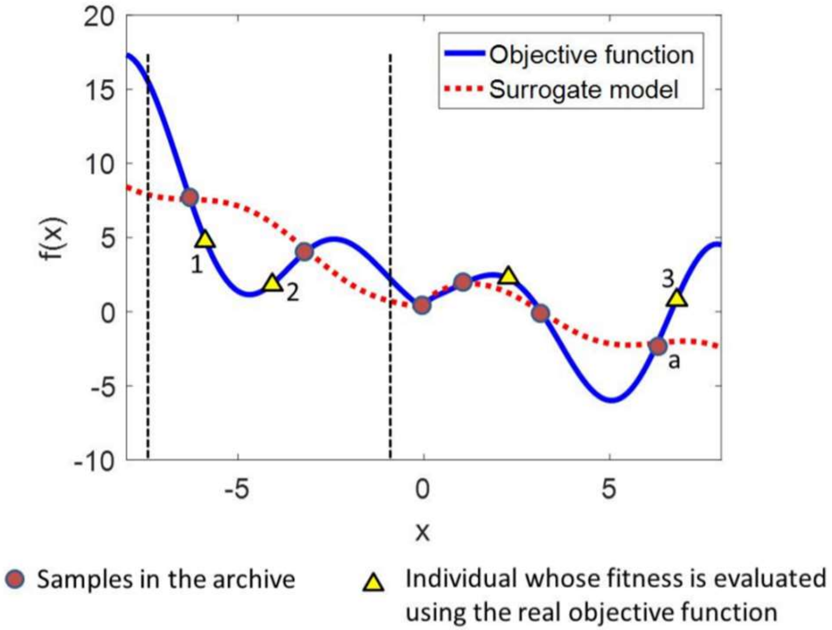
Partial Dependency Plot

- Global method to view few features trends versus predictions
- Partial Dependence shows the marginal impact of a feature on model prediction, holding other features in the model constant
- The derivative of partial dependence describes the impact of a feature
- PDPs can show the type of relationship between the target and a feature



Global Surrogate Models

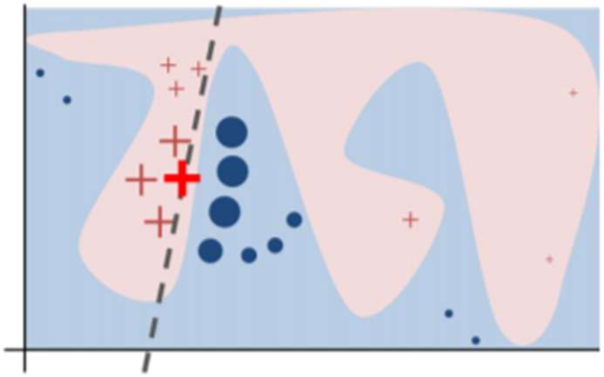
- Interpretable model is trained to approximate as best as possible the predictions of a black box model
- Surrogate model is trained on the black box model predictions
- The interpretation of the surrogate model regard the black box model not the data (surrogate never sees them)



Local Interpretable Model-Agnostic Explanations

LIME

- Local surrogate method
- Surrogate models are interpretable models (like a linear model or decision tree) that are learned on the predictions of the original black box model
- LIME perturbs the dataset to obtain new predictions
- To define a meaningful neighborhood around a point LIME uses an exponential smoothing kernel
- Pipeline:
  - Identify a sample for which interpretation is wanted
  - Perturb the original dataset and obtain predictions
  - Weight the perturbed dataset by the proximity to the sample of interest
  - Train a weighted, interpretable model on the new dataset and new predictions
  - Explain prediction by interpreting the local model.



Ribeiro, et al., ACM 2016

Shapley values

- From cooperative game theory

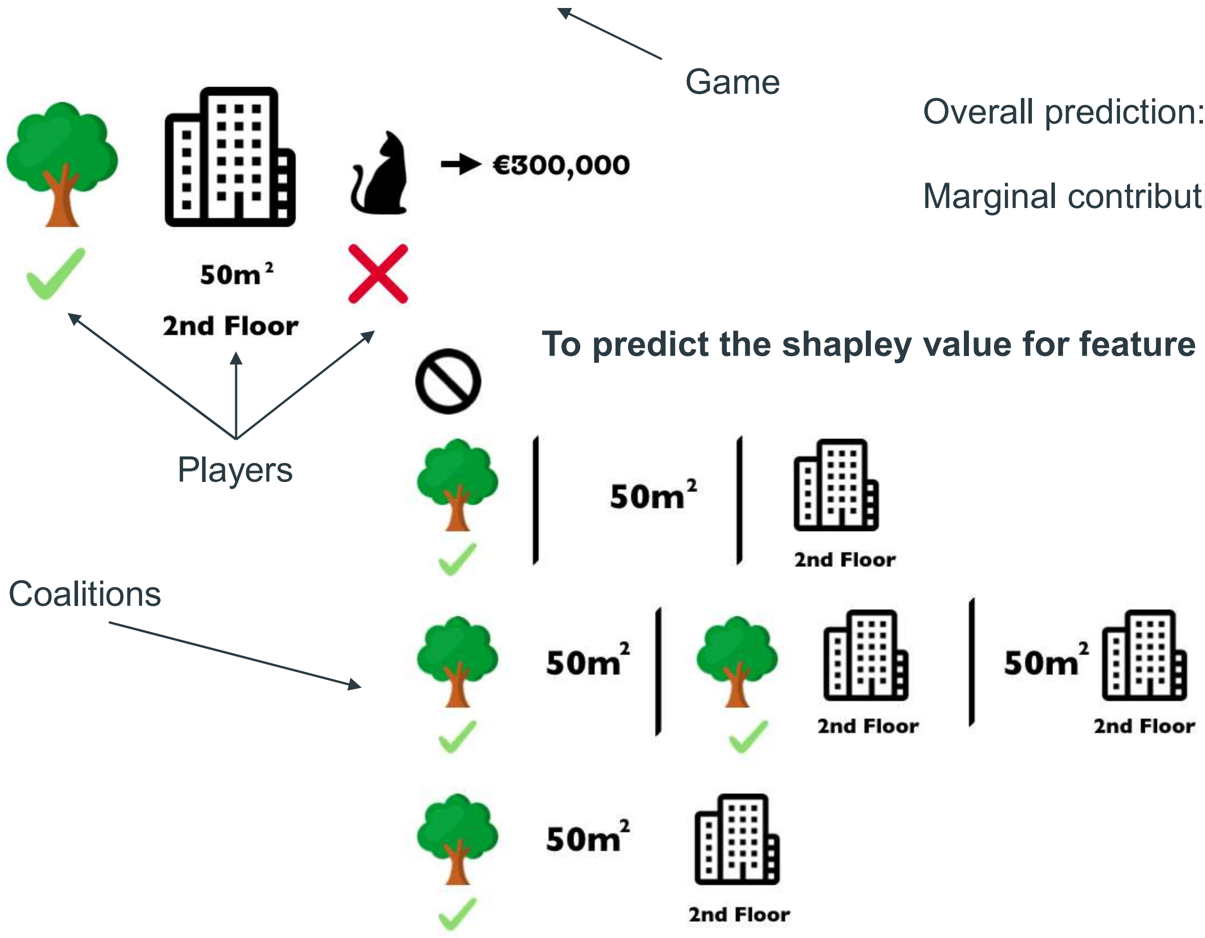
- The Shapley value is a method for assigning payouts to players (features) depending on their contribution towards the total payout (prediction). Players cooperate in a coalition and obtain a certain gain from that cooperation.

- Coalitions are basically combinations of features which are used to estimate the shapley value of a specific feature.

- Given a feature set, find each feature’s marginal contribution to the overall prediction (expected value of the model)
  - The **marginal contribution** would mean how much each feature *forces* the prediction to move away from that baseline.
    - › The marginal contribution is calculated by computing the predicted value with and without the feature value currently being considered and take the difference to get the marginal contribution.
  - Finally, the Shapley value is calculated by averaging the marginal contribution of the feature value across all such possible feature subsets (called coalitions) inside the feature set where the feature participates.
  - To run a model “with and without a feature”: feature values are replaced by random values from the dataset to get a prediction from the machine learning model, or use the expected value for such a feature.

Shapley values step by step

Prediction of apartment price



Overall prediction: The average prediction for all apartments is 310,000€

Marginal contribution: 10,000€ (31 k – 30 k)

To predict the shapley value for feature ‘cat-forbidden’:

1. Replace feature values not in coalition with random numbers

2. Get predictions

3. The Shapley value is the (weighted) average of marginal contributions

SHapley Additive exPlanations

SHAP

- The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction
- Shapley value explanation is represented as an additive feature attribution method, a linear model
- KernelSHAP is a kernel-based estimation approach for Shapley values inspired by local surrogate models (~ LIME)
  - Sample coalitions
  - Get prediction for each coalitions
  - Compute the weight for each coalitions with the SHAP kernel
  - Fit weighted linear model (weight according to the size of the coalitions)
  - Return Shapley values (the coefficients from the linear model)
- TreeSHAP, an efficient estimation approach for tree-based models
  - The Shapley values of a tree ensemble is the (weighted) average of the Shapley values of the individual trees
- In non-linear functions the order in which features are introduced matters. The SHAP values result from averaging over all possible orderings.
- For global importance, average the absolute Shapley values per feature across the data
- For really complex ensemble models there is a high-speed implementation of the algorithm

TREPAN

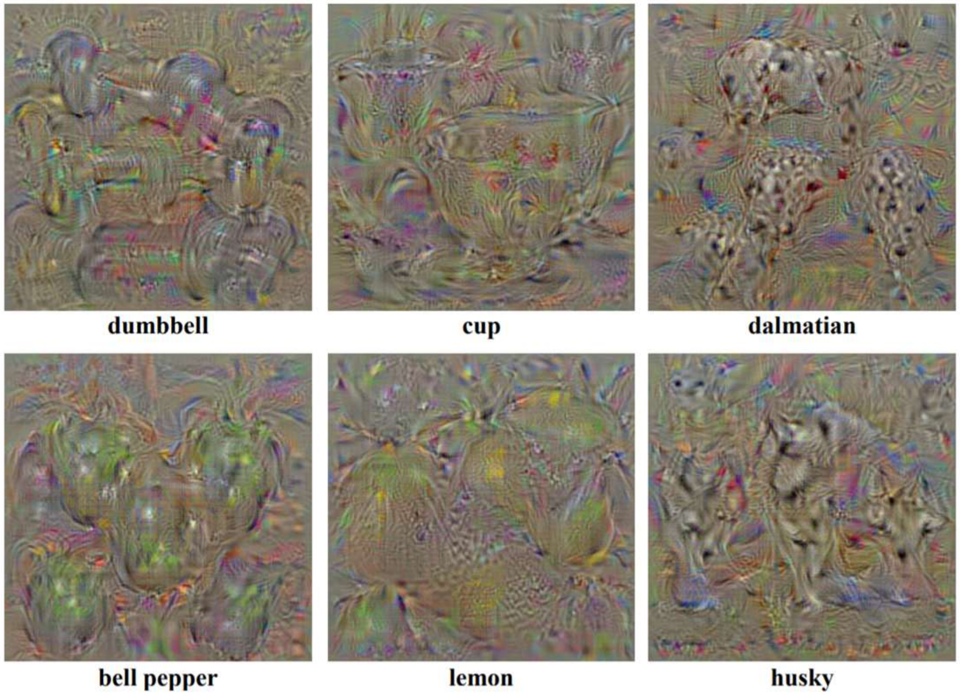
- Decision tree based method
- It has a queue of leaves to expand into new trees
- Each node in the queue consists of:
  - A subset of the **training examples**: instances that reach the node
  - Another set of instances (**query instances**) used with the training example to select the splitting test/class label
  - A set of **constraints**: the conditions that instances must satisfy in order to reach the node; this information is used when drawing a set of query instances for a newly created node.
- For the process of expanding a node, a splitting test is selected for the node, and a child is created for each outcome of the test.
- Each child is either made a leaf of the tree or put into the queue for future expansion.

Interpretation techniques for Neural Network

Deep Neural Network Interpretation Techniques

Backpropagation

- Calculate the gradient, or its variants, of a particular output with respect to the input using back-propagation to derive the contribution of features
- Large gradient for more relevant features



dumbbell cup dalmatian

bell pepper lemon husky

Simonyan, Vedaldi, Zisserman, 2014, https://arxiv.org/pdf/1312.6034.pdf

Deep Neural Network Interpretation Techniques

Activation maximization

- Study what each neuron has learned to detect
- Synthesize an input (e.g. an image) that highly activates a neuron

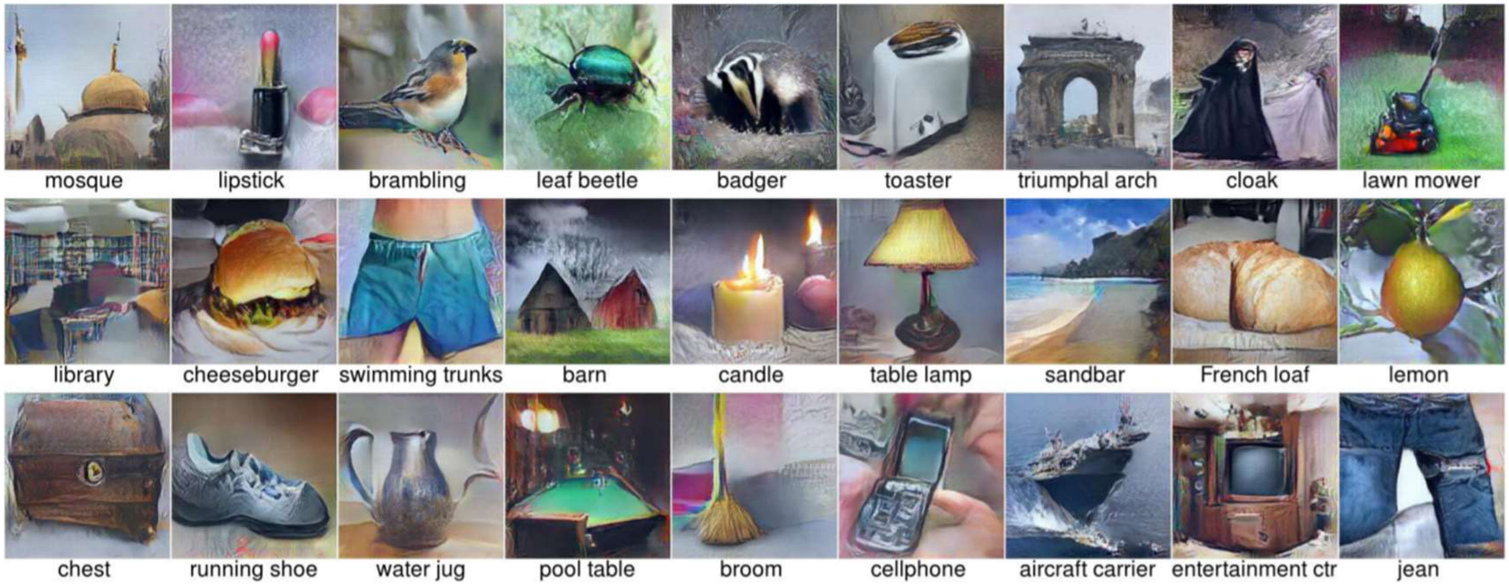


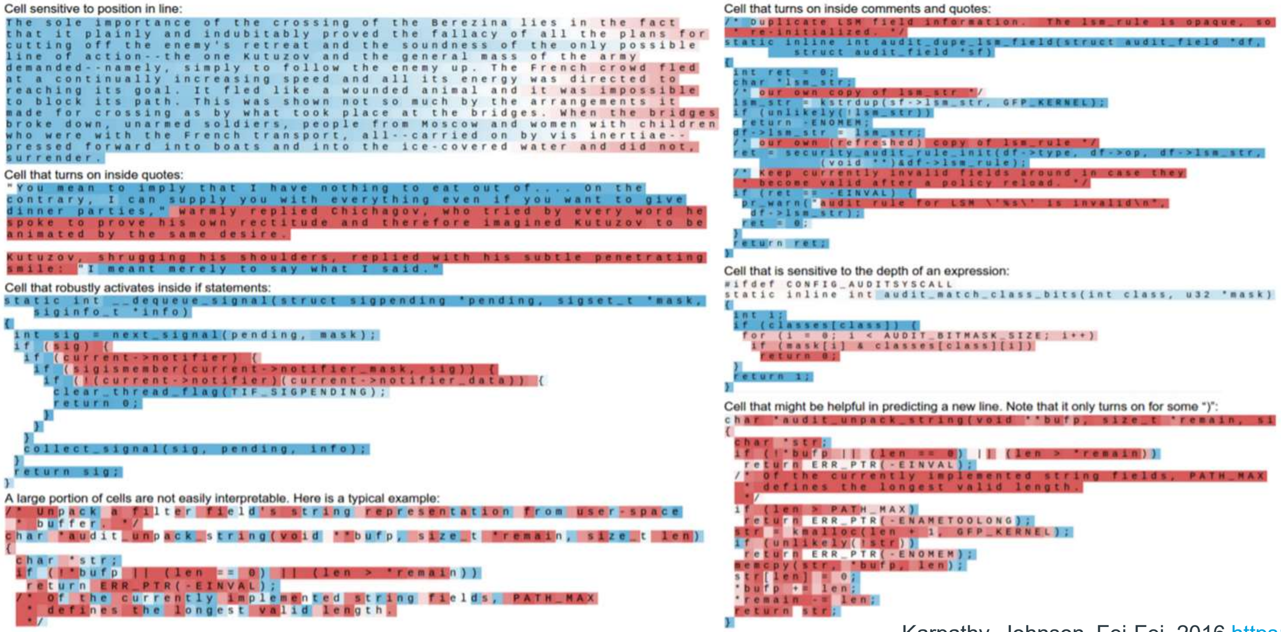
Figure 1: Images synthesized from scratch to highly activate output neurons in the CaffeNet deep neural network, which has learned to classify different types of ImageNet images.

Nguyen et al, NIPS 2016

Interpreting RNN

Which part of text maximize a unit activation

- Approach similar to CNN interpretation
- Representations of the last hidden layer of RNN and study the function of different units at that layer, by analyzing the real input tokens that maximally activate a unit.
- CNN is able to capture complex language characteristics, e.g., syntax, semantics and long-term dependencies.



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
    siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
      struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
          (void **)&df->lsm_rule);
  /* Keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \'%s\' is invalid\n",
      df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

Cell that might be helpful in predicting a new line. Note that it only turns on for some ")":

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
  if (len > PATH_MAX)
    return ERR_PTR(-ENAMETOOLONG);
  str = kmalloc(len + 1, GFP_KERNEL);
  if (unlikely(!str))
    return ERR_PTR(-ENOMEM);
  memcpy(str, *bufp, len);
  str[len] = 0;
  *bufp += len;
  *remain -= len;
  return str;
}
```

Karpathy, Johnson, Fei-Fei, 2016 https://arxiv.org/pdf/1506.02078.pdf

Deep Neural Network Interpretation Techniques

Mask perturbation

- Perturbation of the input to learn a perturbation mask which explicitly preserves the contribution values of each feature
  - Use regularizations
- Masking salient parts of the input image, to yield attribution scores for an input
- Care must be put on the perturbation technique, not to add artifacts

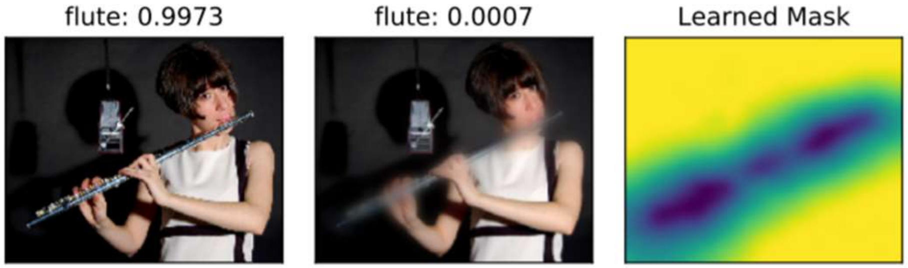


Figure 1. An example of a mask learned (right) by blurring an image (middle) to suppress the softmax probability of its target class (left: original image; softmax scores above images).

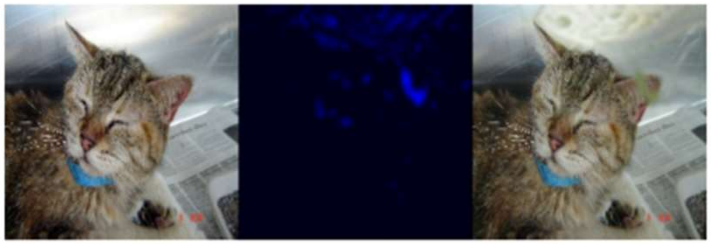


Figure 3: The adversarial mask introduces very small perturbations, but can completely alter the classifier's predictions. From left to right: an image which is correctly recognised by the classifier with a high confidence as a "tabby cat"; a generated adversarial mask; an original image after application of the mask that is no longer recognised as a "tabby cat".

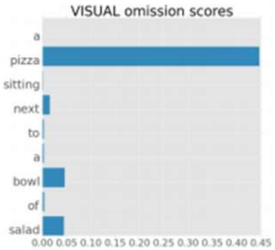
Fong and Vedaldi, 2018 https://arxiv.org/pdf/1704.03296.pdf

Dabkowski and Gal, 2017 https://arxiv.org/pdf/1705.07857.pdf

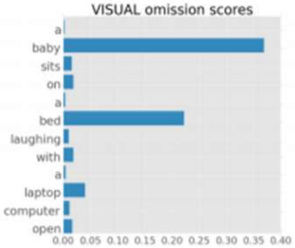
Interpreting RNN

RNN is able to learn linguistic structures

- Saliency was measured omitting the word
    - “the black dog is running” -> “the dog is running”
- Lexical categories and grammatical functions that carry semantic information are learnt by the network
- The same input token is treated differently depending on its grammatical functions in the sentence
- Some hidden units could carry the activation values over to subsequent time steps (long term dependencies)



a **pizza** sitting next to a bowl of salad



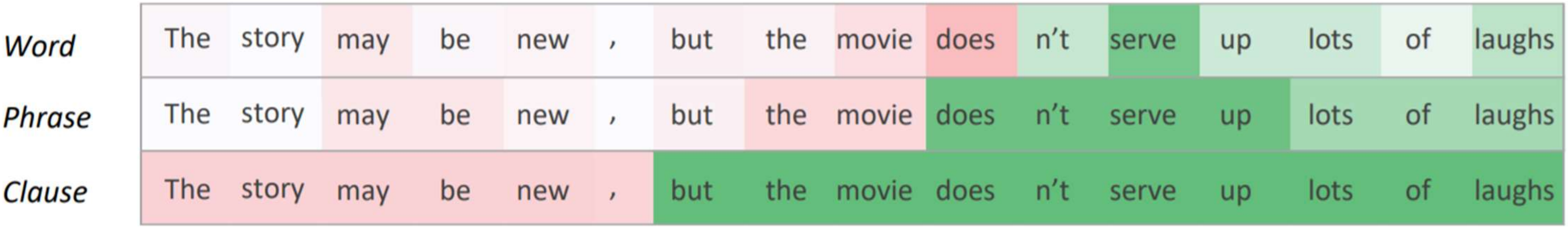
a **baby** sits on a bed laughing with a laptop computer open

Kádár, Chrupała, Alishahi , 2016 https://arxiv.org/pdf/1602.08952.pdf

Interpreting RNN

Additive decomposition

- RNN prediction is decomposed into additive contribution of each word in the input text, allowing quantification of the contribution of each individual word to a RNN prediction
- It is also possible to obtain a phrase-level attribution



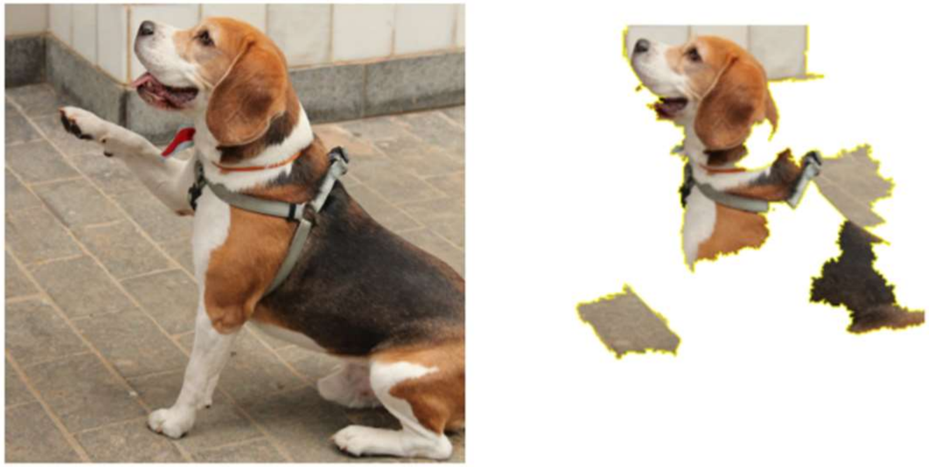
Word	The	story	may	be	new	,	but	the	movie	does	n't	serve	up	lots	of	laughs
Phrase	The	story	may	be	new	,	but	the	movie	does	n't	serve	up	lots	of	laughs
Clause	The	story	may	be	new	,	but	the	movie	does	n't	serve	up	lots	of	laughs

Du et al., WWW 2019, https://arxiv.org/pdf/1903.11245.pdf

Local interpretation of CNN

Local approximation based explanation

- Modelling the surrounding of predictions using a linear model
  - Sample the feature space in the neighborhood of the instance to constitute an additional training set.
  - Train a white box model (eg Lasso or decision rule)
  - Predict the original model using the weights of the model



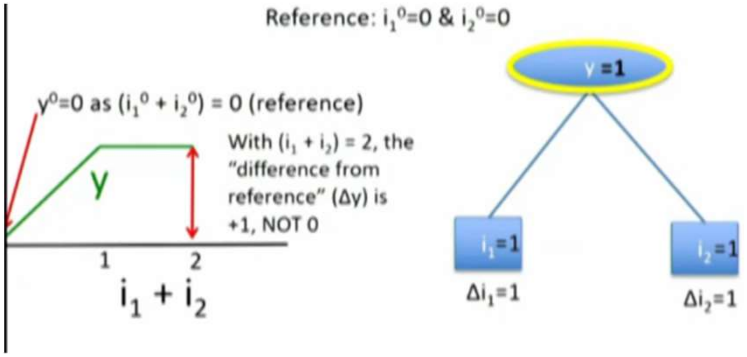
(a) Original image (b) Anchor for "beagle"

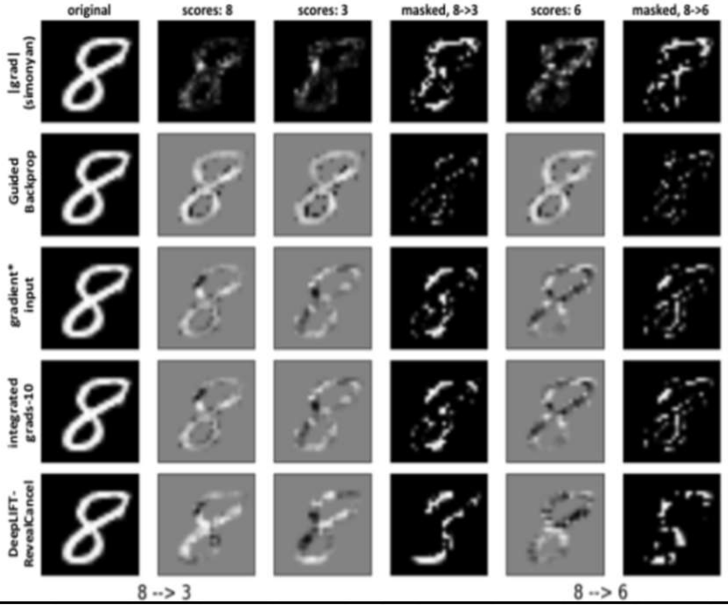
Ribeiro, Singh, Guestrin, 2016, AAAI 2018 https://arxiv.org/pdf/1602.04938.pdf https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

Deep Learning Important FeaTures

DeepLift

- Local interpretation
- Difference from reference (or neutral) input, eg i1=i2=0
  - Propagates importance even when gradient is 0 (saturation)
- Choice of reference is important
  - Use reference as distribution (eg insignificant input)


Reference: i_1^0=0 & i_2^0=0
y^0=0 as (i_1^0 + i_2^0) = 0 (reference)
y
With (i_1 + i_2) = 2, the "difference from reference" (Δy) is +1, NOT 0
1
2
$i_1 + i_2$
y =1
i_1=1
i_2=1
Δi_1=1
Δi_2=1



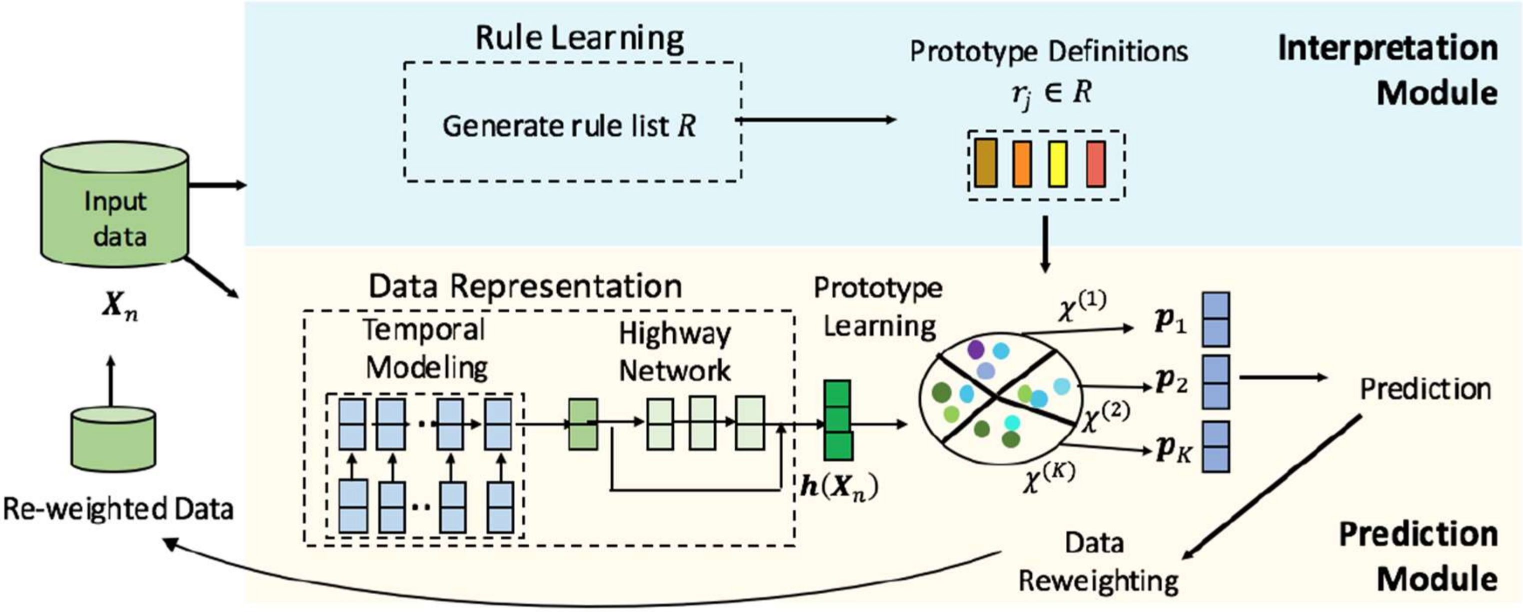
DeepLIFT with the RevealCancel rule better identifies pixels to convert one digit to another.

Shrikumar, Greenside, Kundaje ICML 2017

Prototype Learning via Rule Learning

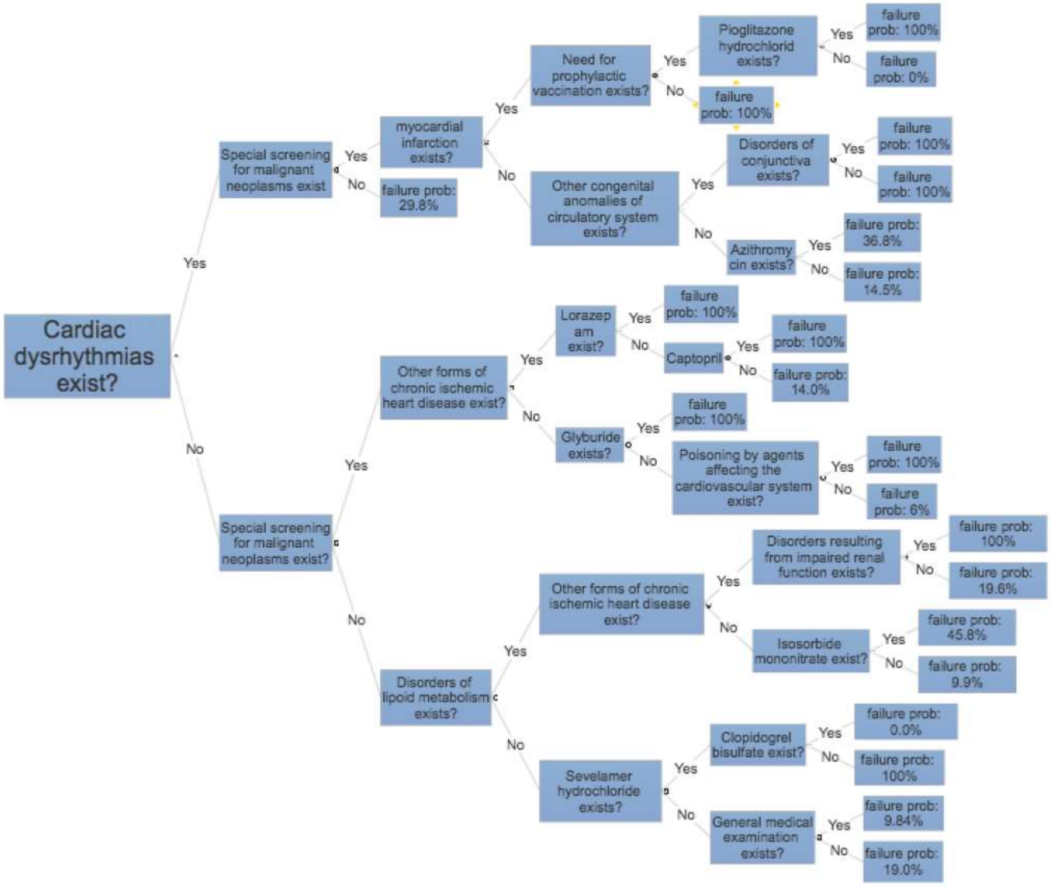
PEARL

- Prototype learning: observations are classified based on their proximity to a prototype point in the dataset
- Iteratively learn decision rules, via a data reweighing procedure using prototypes, and then update prototypes via neural networks with learned rules



Fu et al., ACM-BCB 2019

Example of decision rules and predictions



Cardiac dysrhythmias exist?

Yes

Special screening for malignant neoplasms exist

Yes

myocardial infarction exists?

Yes

Need for prophylactic vaccination exists?

Yes

Pioglitazone hydrochlorid exists?

Yes

failure prob: 100%

No

failure prob: 0%

No

failure prob: 100%

No

Other congenital anomalies of circulatory system exists?

Yes

Disorders of conjunctiva exists?

Yes

failure prob: 100%

No

failure prob: 100%

No

Azithromycin exists?

Yes

failure prob: 36.8%

No

failure prob: 14.5%

No

failure prob: 29.8%

No

Special screening for malignant neoplasms exist?

Yes

Other forms of chronic ischemic heart disease exist?

Yes

Lorazepam exist?

Yes

failure prob: 100%

No

Captopril

Yes

failure prob: 100%

No

failure prob: 14.0%

No

Glyburide exists?

Yes

failure prob: 100%

No

Poisoning by agents affecting the cardiovascular system exist?

Yes

failure prob: 100%

No

failure prob: 6%

No

Disorders of lipoid metabolism exists?

Yes

Other forms of chronic ischemic heart disease exist?

Yes

Disorders resulting from impaired renal function exists?

Yes

failure prob: 100%

No

failure prob: 19.6%

No

Isosorbide mononitrate exist?

Yes

failure prob: 45.8%

No

failure prob: 9.9%

No

Sevelamer hydrochloride exists?

Yes

Clopidogrel bisulfate exist?

Yes

failure prob: 0.0%

No

failure prob: 100%

No

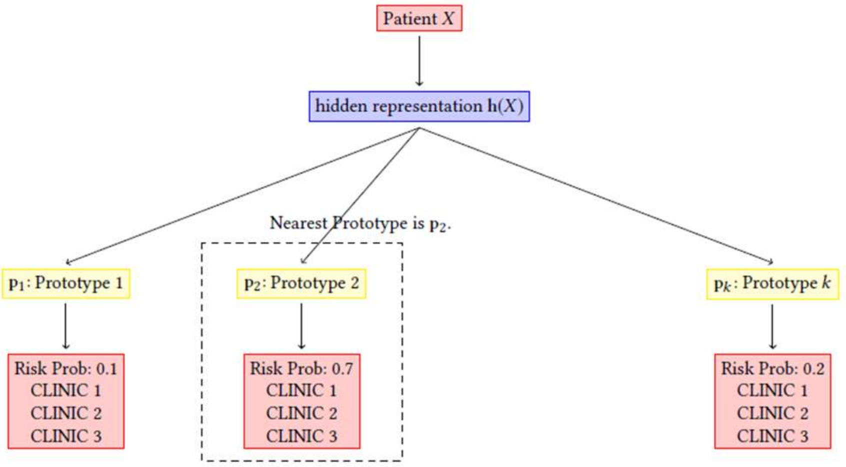
General medical examination exists?

Yes

failure prob: 9.84%

No

failure prob: 19.0%



Patient X

hidden representation $\mathbf{h}(X)$

Nearest Prototype is $\mathbf{p}_2$.

$\mathbf{p}_1$: Prototype 1

Risk Prob: 0.1
CLINIC 1
CLINIC 2
CLINIC 3

$\mathbf{p}_2$: Prototype 2

Risk Prob: 0.7
CLINIC 1
CLINIC 2
CLINIC 3

$\mathbf{p}_k$: Prototype k

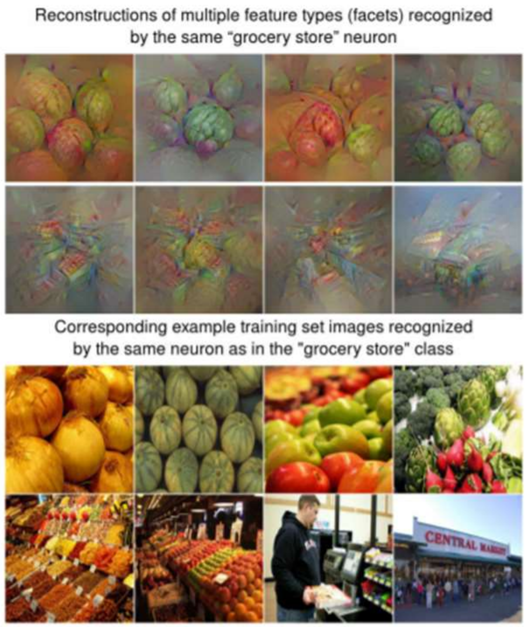
Risk Prob: 0.2
CLINIC 1
CLINIC 2
CLINIC 3

Fu et al., ACM-BCB 2019

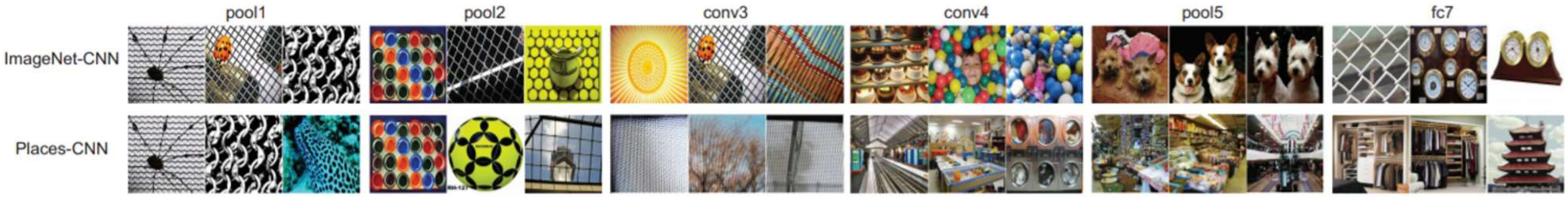
Interpreting CNN

Insights from CNN interpretation

- CNN representations are learnt at different levels of abstraction, transiting from general at lower levels to task-specific to the last layers
- A neuron could respond to different images that are related to a semantic concept, revealing the multifaceted nature of neurons
- Objects can be described using part-based representations and these parts can be shared across different categories


Reconstructions of multiple feature types (facets) recognized by the same "grocery store" neuron
Corresponding example training set images recognized by the same neuron as in the "grocery store" class
CENTRAL MARKET

Nguyen, Yosinski, Clune, ICML 2016 http://www.evolvingai.org/files/mfv_icml_workshop_16.pdf

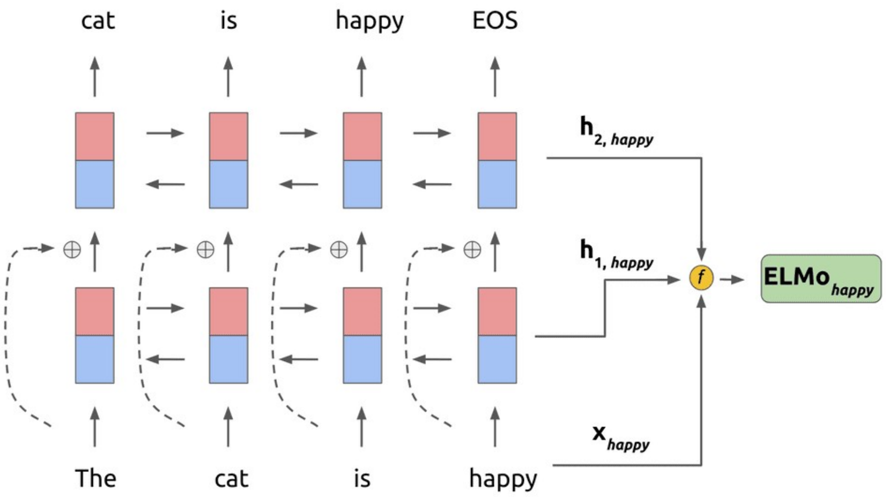


Zhou et al., ICLR 2015, https://arxiv.org/pdf/1412.6856.pdf

Interpreting RNN

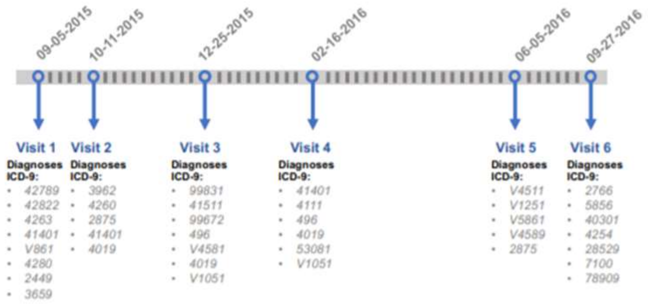
Complex architectures learn hierarchical language models

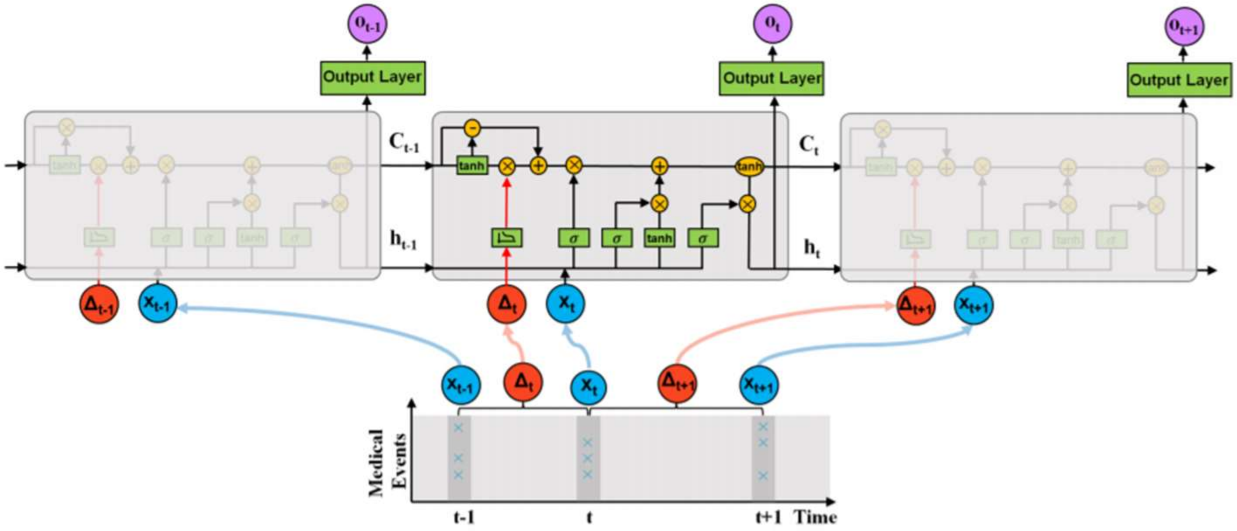
- The complex architecture of ELMo can capture very rich language representation in a hierarchical fashion
- Higher-level LSTM states disambiguate the meanings of words by utilizing their context to capture word meaning
  - Helpful on word sense disambiguation tasks
- Lower level states model aspects of syntax
  - Can be used for part-of-speech tagging tasks

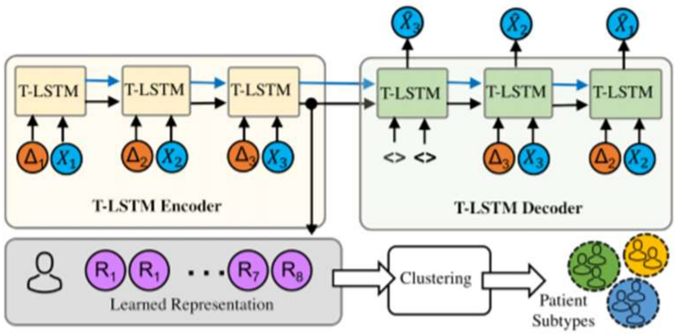


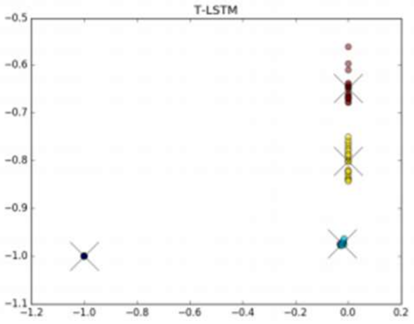
Peters et al, NAACL 2018 https://arxiv.org/pdf/1802.05365.pdf

Patient subtyping via Time-Aware LSTM Networks









(a) T-LSTM Auto-Encoder

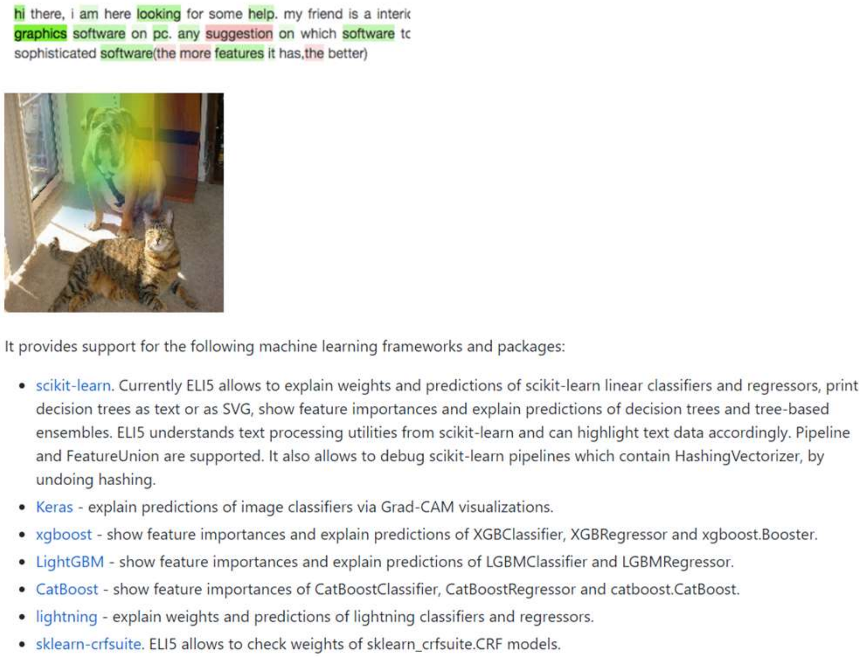
Baytas et al, KDD 2017

Interpretation frameworks

Interpretation wrappers

Eli5

- https://github.com/TeamHG-Memex/eli5
- For tree-based models it wraps feature importance (default: gain)
- Prediction can be calculated as the sum of the feature contributions + the “bias”



hi there, i am here looking for some help. my friend is a interic
graphics software on pc. any suggestion on which software tc
sophisticated software(the more features it has,the better)

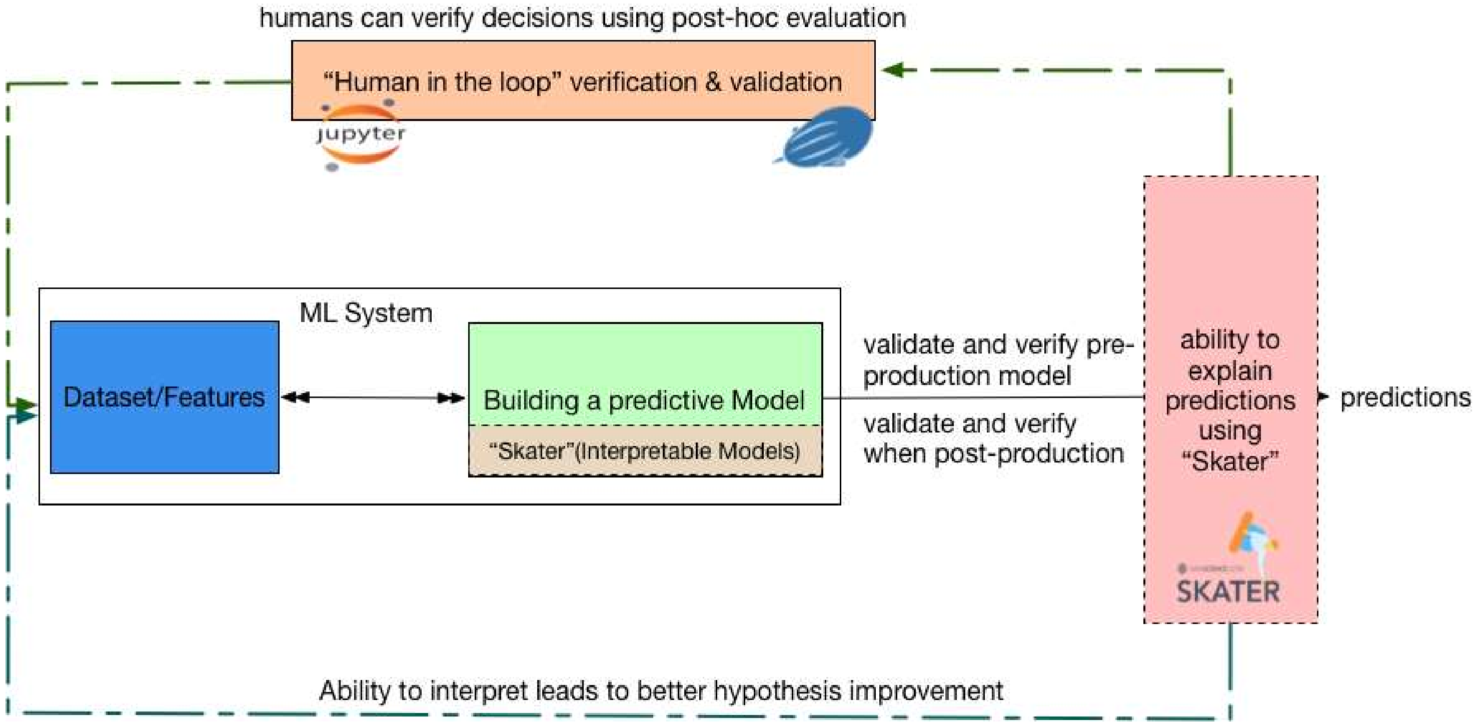
It provides support for the following machine learning frameworks and packages:

- scikit-learn. Currently ELI5 allows to explain weights and predictions of scikit-learn linear classifiers and regressors, print decision trees as text or as SVG, show feature importances and explain predictions of decision trees and tree-based ensembles. ELI5 understands text processing utilities from scikit-learn and can highlight text data accordingly. Pipeline and FeatureUnion are supported. It also allows to debug scikit-learn pipelines which contain HashingVectorizer, by undoing hashing.
- Keras - explain predictions of image classifiers via Grad-CAM visualizations.
- xgboost - show feature importances and explain predictions of XGBClassifier, XGBRegressor and xgboost.Booster.
- LightGBM - show feature importances and explain predictions of LGBMClassifier and LGBMRegressor.
- CatBoost - show feature importances of CatBoostClassifier, CatBoostRegressor and catboost.CatBoost.
- lightning - explain weights and predictions of lightning classifiers and regressors.
- sklearn-crfsuite. ELI5 allows to check weights of sklearn_crfsuite.CRF models.

Interpretation wrappers

Skater

- https://github.com/oracle/Skater
- Started off as a fork of LIME


humans can verify decisions using post-hoc evaluation
“Human in the loop” verification & validation
jupyter
ML System
Dataset/Features
Building a predictive Model
“Skater”(Interpretable Models)
validate and verify pre-production model
validate and verify when post-production
ability to explain predictions using “Skater”
SKATER
predictions
Ability to interpret leads to better hypothesis improvement

Interpretation wrappers

Skater suite of model interpretation techniques

- Global interpretation
  - › Feature importance
  - › Partial dependency plot
- Local interpretation
  - › LIME (local interpretable model explanation)
  - › DNN
    - » Layer-wise relevance propagation
    - » Occlusion
    - » Integrated Gradient
- Global and local interpretation
  - › Scalable Bayesian Rule Lists
  - › Tree Surrogates

Hands-on model interpretation

Github: https://github.com/paganilucia/ODSC_2020_Milan
Dataset: http://www.emrbots.org/Downloads2222694716.html

Business problem

Identify subjects likely to be included in healthy arm of clinical study

- Data: synthetic Electronic Health Record
- Downloadable at http://www.emrbots.org/
- Very similar to original EHR

Data exploration

- Demographic data
  - [PatientID] - a unique ID representing a patient.
  - [PatientGender] - Male/Female.
  - [PatientDateOfBirth] - Date Of Birth.
  - [PatientRace] - African American, Asian, White.
  - [PatientMaritalStatus] - Single, Married, Divorced, Separated, Widowed.
  - [PatientLanguage] - English, Icelandic, Spanish.
  - [PatientPopulationPercentageBelowPoverty] - given in %.
- Admission data
  - [PatientID] - a unique ID representing a patient.
  - [AdmissionID] - an admission ID for the patient.
  - [AdmissionStartDate] - start date.
  - [AdmissionEndDate] - end date.

Data Exploration

- Diagnoses data
  - [PatientID] - a unique ID representing a patient.
  - [AdmissionID] - an admission ID for the patient.
  - [PrimaryDiagnosisCode] - ICD10 code for admission's primary diagnosis.
  - [PrimaryDiagnosisDescription] - admission's primary diagnosis description.
- Primary Diagnosis Code is given in ICD10 code

ICD10 code

Chapter	Block	Title
I	A00–B99	Certain infectious and parasitic diseases
II	C00–D48	Neoplasms
III	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00–E90	Endocrine, nutritional and metabolic diseases
V	F00–F99	Mental and behavioural disorders
VI	G00–G99	Diseases of the nervous system
VII	H00–H59	Diseases of the eye and adnexa
VIII	H60–H95	Diseases of the ear and mastoid process
IX	I00–I99	Diseases of the circulatory system
X	J00–J99	Diseases of the respiratory system
XI	K00–K93	Diseases of the digestive system
XII	L00–L99	Diseases of the skin and subcutaneous tissue
XIII	M00–M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00–N99	Diseases of the genitourinary system
XV	O00–O99	Pregnancy, childbirth and the puerperium
XVI	P00–P96	Certain conditions originating in the perinatal period
XVII	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00–T98	Injury, poisoning and certain other consequences of external causes
XX	V01–Y98	External causes of morbidity and mortality
XXI	Z00–Z99	Factors influencing health status and contact with health services
XXII	U00–U99	Codes for special purposes

Data preprocessing

- Joined the tables
- Obtained age in years
- Obtained admitted days and whether the admission was long
- Binarized gender
- Converted categorical features in classes
- Obtained features for every subject
  - Age at (second) last admission
  - Total number of admissions (-1)
  - Percentage of long admissions (> 2 days)
  - Past primary diagnosis
  - Gender, race, marital status, language, poverty
- Obtained labels for every subject
  - no_admissions for the next 7 years
- Made BOW on past diagnosis

Let’s dig into the code

https://github.com/paganilucia/ODSC_2020_Milan

Take-home messages

Fear no more complex models!

- Explainable AI is a fast growing field
- Currently many different techniques are available to explain black box models
- Any technique is not good on all models, but all techniques are useful on any model

Model complexity vs interpretability

Can you have it all?

Yes, you can! And you should!

Thank you for your attention!

- Have a good model interpretation!
- And a good lunch!

Resources

- https://christophm.github.io/interpretable-ml-book/
- https://towardsdatascience.com/human-interpretable-machine-learning-part- [1 to 4]
- https://arxiv.org/pdf/1808.00033.pdf
- https://drive.google.com/file/d/1Bw9aBq5VbQY5C5t0RAi9sFGSp3gQFu6V/view
- https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f
- https://community.fico.com/s/explainable-machine-learning-challenge
- http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- https://awesomeopensource.com/project/slundberg/shap