Introduzione alla Explainability
Carlo Batini

Esempio di partenza

- Anche se non sapete nulla di linguaggi programmativi, provate a cercare di capire “cosa” calcolano i seguenti due programmi

Programma 1

- SOMMA = 100 * (100 + 1) / 2
- TERMINA

Programma 2

- SOMMA = 0
- PER I CHE VA DA 1 A 100 ESEGUI
- SOMMA = SOMMA + I
- I = I + 1
- TERMINA

Risposta

l programma 1 calcola con una sola istruzione il valore che corrisponde alla somma dei primi 100 interi. Se non siete convinti, provate a verificare la validità della formula per i primi 5 numeri:

somma = 1 + 2 + 3 + 4 + 5 + 6 = 15 = 30 = 5 *6 / 2

Il secondo programma calcola lo stesso numero, ma in modo più complicato, calcolando prima 0 + 1, poi incrementando la variabile I di 1, aggiungendo a SOMMA il valore 23 e così via, fino a 100.

Commento come spiegazione - 1

Certamente, invece che dover capire da soli il calcolo effettuato dal programma, sarebbe utile che all'inizio dei due programmi comparisse una frase, detta commento, così concepita:

COMMENTO – IL PROGRAMMA CALCOLA LA SOMMA DEI PRIMI CENTO NUMERI

Il precedente commento chiarisce "cosa" fa il programma, non come lo fa. Certe volte, può essere utile conoscere oltre al cosa il come, ad esempio per poter valutare la efficienza dell'algoritmo, misurata dal numero di istruzioni eseguite. Possiamo perciò aggiungere

Commento come spiegazione - 2

Al Programma 1 il commento

- LA SOMMA E' CALCOLATA IN BASE AD UNA SEMPLICE FORMULA MATEMATICA CHE LEGA IL NUMERO N ALLA SOMMA DEI PRIMI N NUMERI INTERI

Al programma 2 il commento

- LA SOMMA E' CALCOLATA SOMMANDO AL VALORE 0 SUCCESSIVAMENTE I VALORI 1, 2, …, FINO A 100.

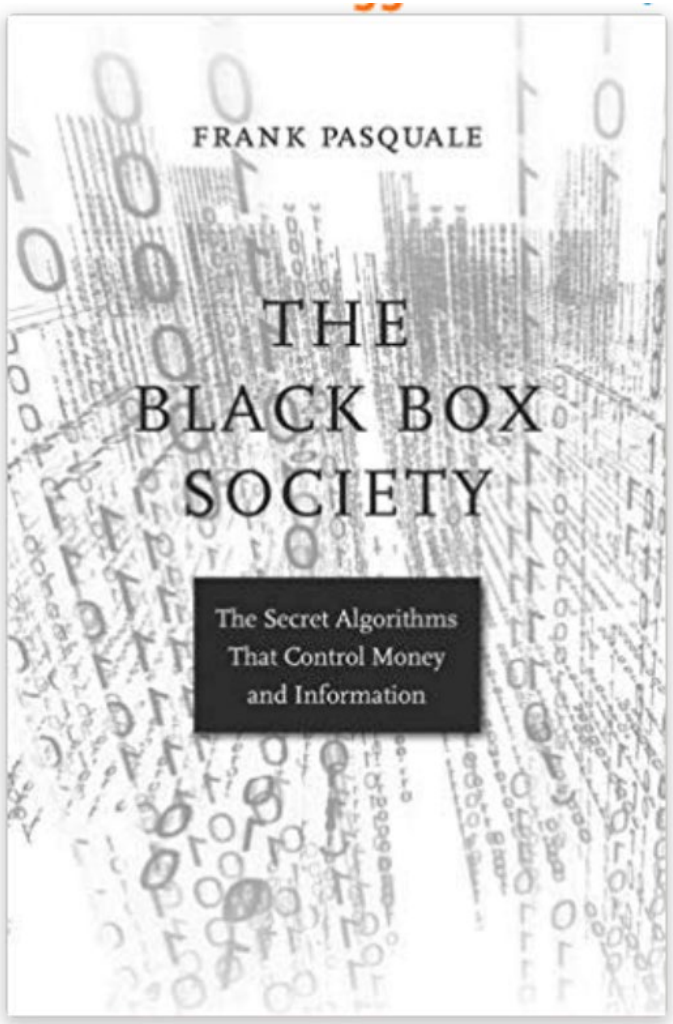
Dai programmi al machine learning

- E’ chiaro che il secondo programma esegue più istruzioni del primo; il primo programma ha anche la caratteristica di essere scalabile, intendendo che il numero di istruzioni eseguite per sommare i primi 1.000, 10.000, ecc . numeri è sempre lo stesso, mentre invece il numero di istruzioni cresce linearmente nel secondo programma.
- Il problema di rendere gli algoritmi comprensibili agli umani, già presente nella informatica da quando esiste il software, diventa ancor più critico nell’epoca del Machine learning, perchè ora l’algoritmo che prima veniva concepito e prodotto da esseri umani, ora è autonomamente prodotto dalla stessa tecnica di apprendimento, come già osservato nella sezione precedente con riferimento alla equità.

La Black Box society

- Il grande rischio che corriamo è quello che [Pasquale 2015] chiama la “black box society”, una società in cui le decisioni sono prese basandosi su analisi e modelli previsionali i cui meccanismi di funzionamento sono noti solo a pochi, e certe volte neanche ad essi, ma solo a chi li ha concepiti, soggetti che per ragioni di concorrenza o sicurezza, non intendono o non vogliono condividerli.

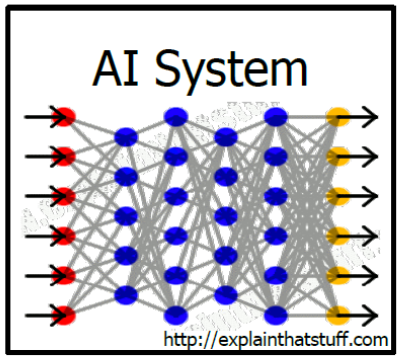
The black box society



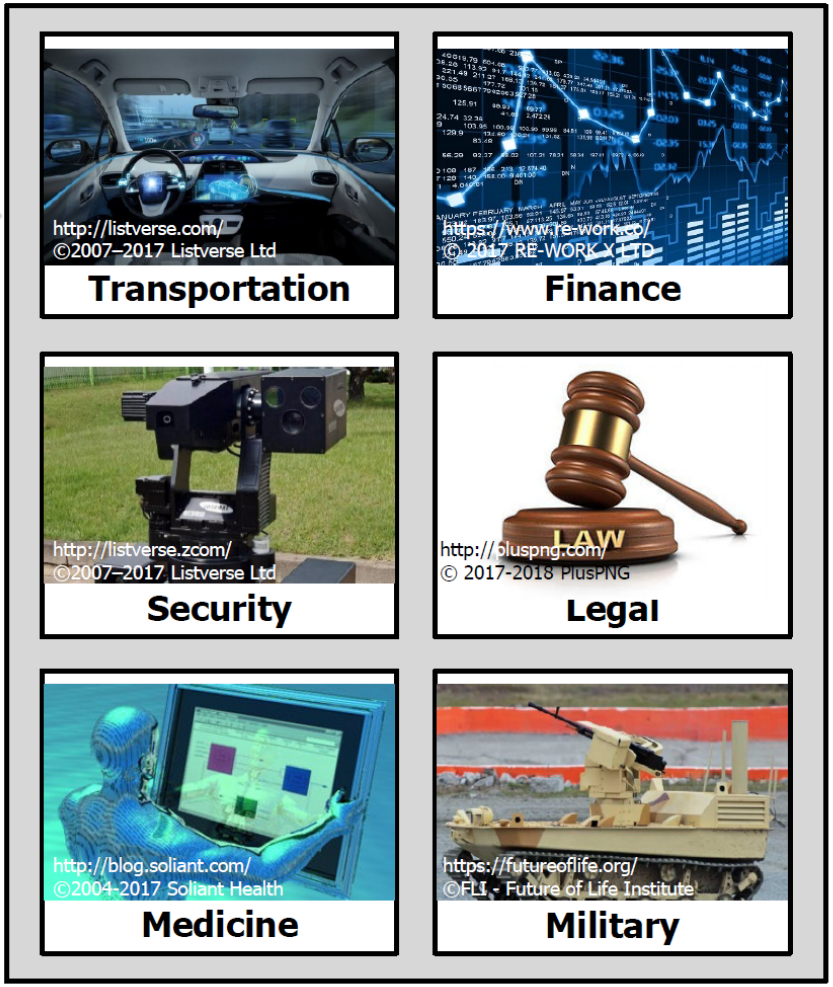
Opacity and bias

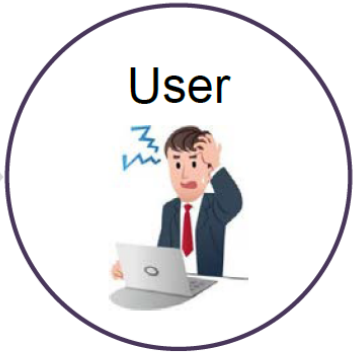
- Black boxes map user features into a class or a score without explaining why, because the decision model is not comprehensible to stakeholders, even to expert data scientists.
- This is worrying not only for the lack of transparency, but also for the possible biases hidden in the algorithms.

Scopi della explainability



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

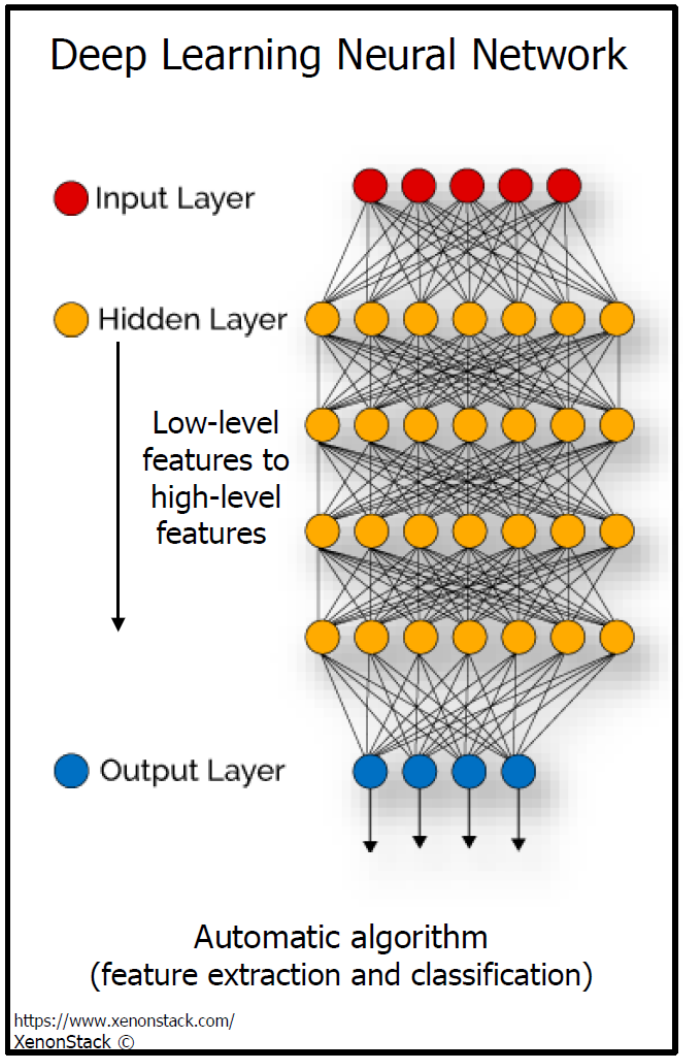


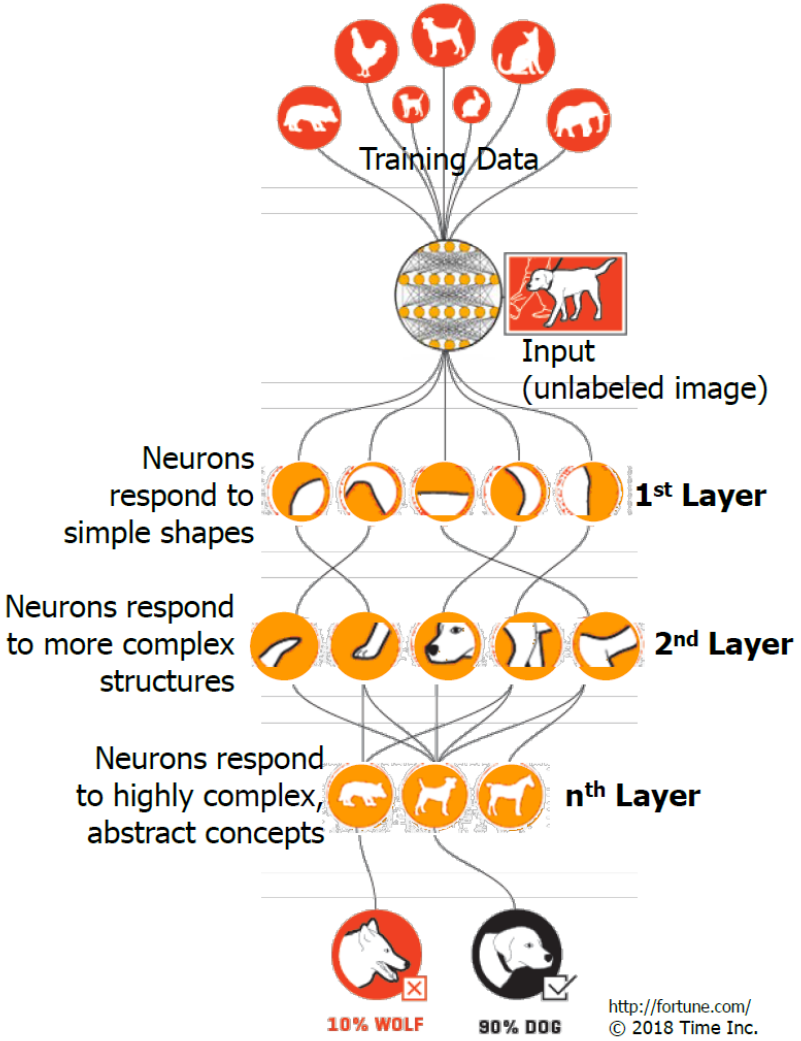


- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users

Esempio di tecnica di deep learning
la tecnica è molto potente ma opaca
nello spiegare la propria logica e strategia



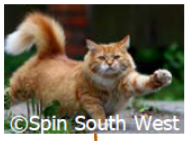


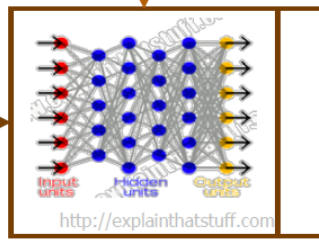
Processo di apprendimento oggi e in futuro in forma “explainable”

Today

Training Data

Learning Process





Learned Function

This is a cat
(p = .93)

Output

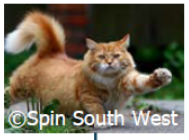
User with a Task

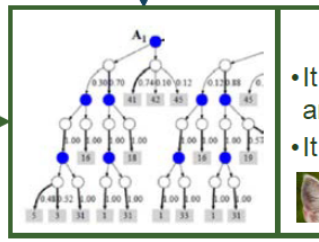
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow

Training Data

New Learning Process





Explainable Model

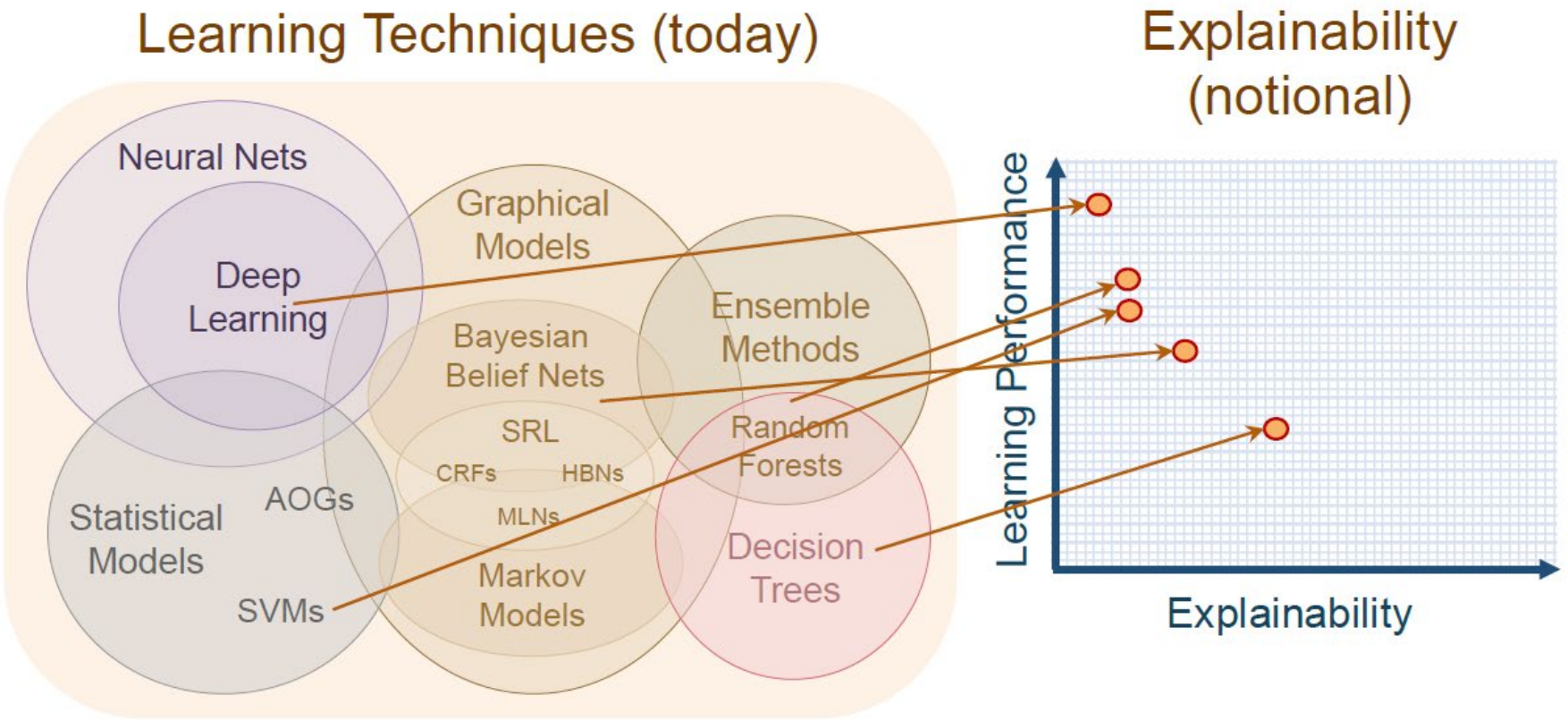
This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

Explanation Interface

User with a Task

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

Oggi

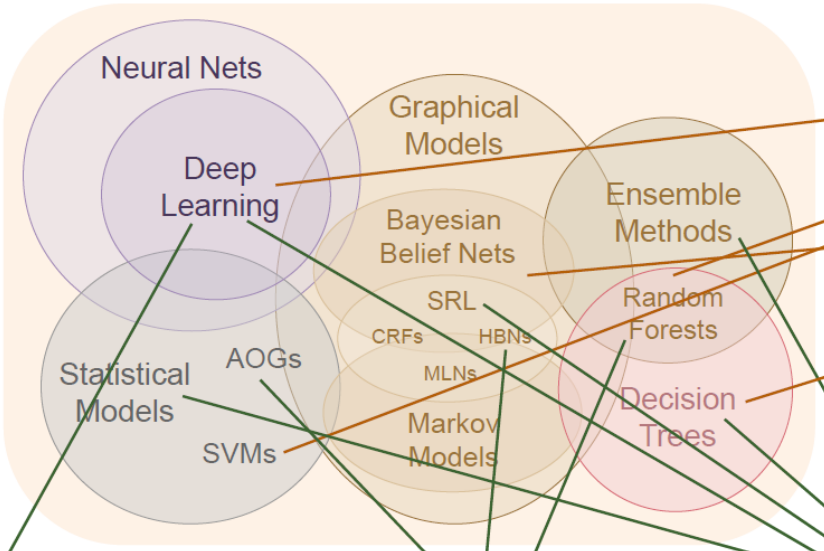

Learning Techniques (today)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
SRL
CRFs
HBNs
MLNs
Markov Models
Ensemble Methods
Random Forests
Decision Trees
Statistical Models
AOGs
SVMs
Explainability (notional)
Learning Performance
Explainability

Domani..

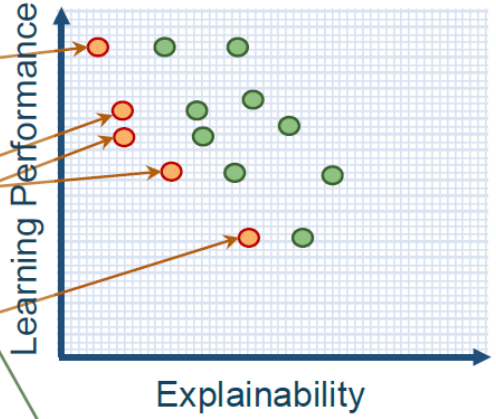
New Approach

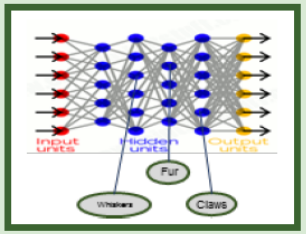
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



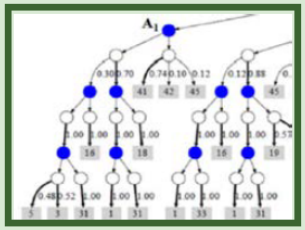
Explainability (notional)





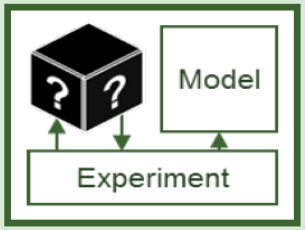
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



Model Induction

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

Tradeoffs che coinvolgono la explainability

Accuracy vs intelligibility

- In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy. This tradeoff sometimes limits the accuracy of models that can be applied in mission-critical applications such as healthcare where being able to understand, validate, edit, and trust a learned model is important.

Accuracy vs intelligibility

- We discuss two case studies where high-performance generalized additive models with pairwise interactions (GA2Ms) are applied to real healthcare problems yielding intelligible models with state-of-the-art accuracy. In the pneumonia risk prediction case study, the intelligible model uncovers surprising patterns in the data that previously had prevented complex learned models from being fielded in this domain, but because it is intelligible and modular allows these patterns to be recognized and removed.
- In the 30 day hospital readmission case study, we show that the same methods scale to large datasets containing hundreds of thousands of patients and thousands of attributes while remaining intelligible and providing accuracy comparable to the best (unintelligible) machine learning methods