Accessible artificial intelligence for data science

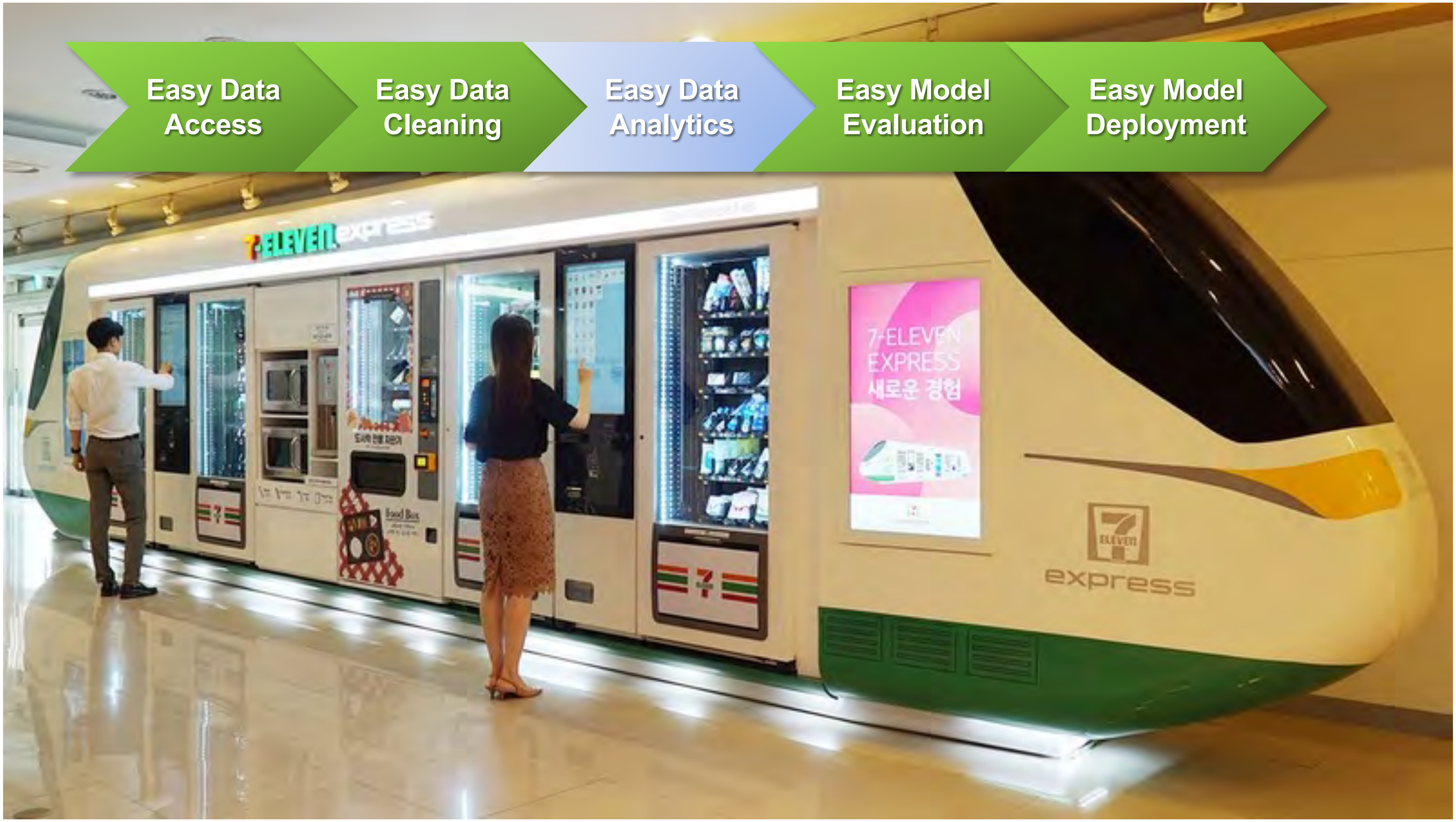
Jason H. Moore, PhD, FACMI, FIAHSI, FASA
Chair, Department of Computational Biomedicine
Director, Center for Artificial Intelligence Research and Education
Cedars-Sinai Medical Center
Los Angeles, CA

jason.moore@csmc.edu

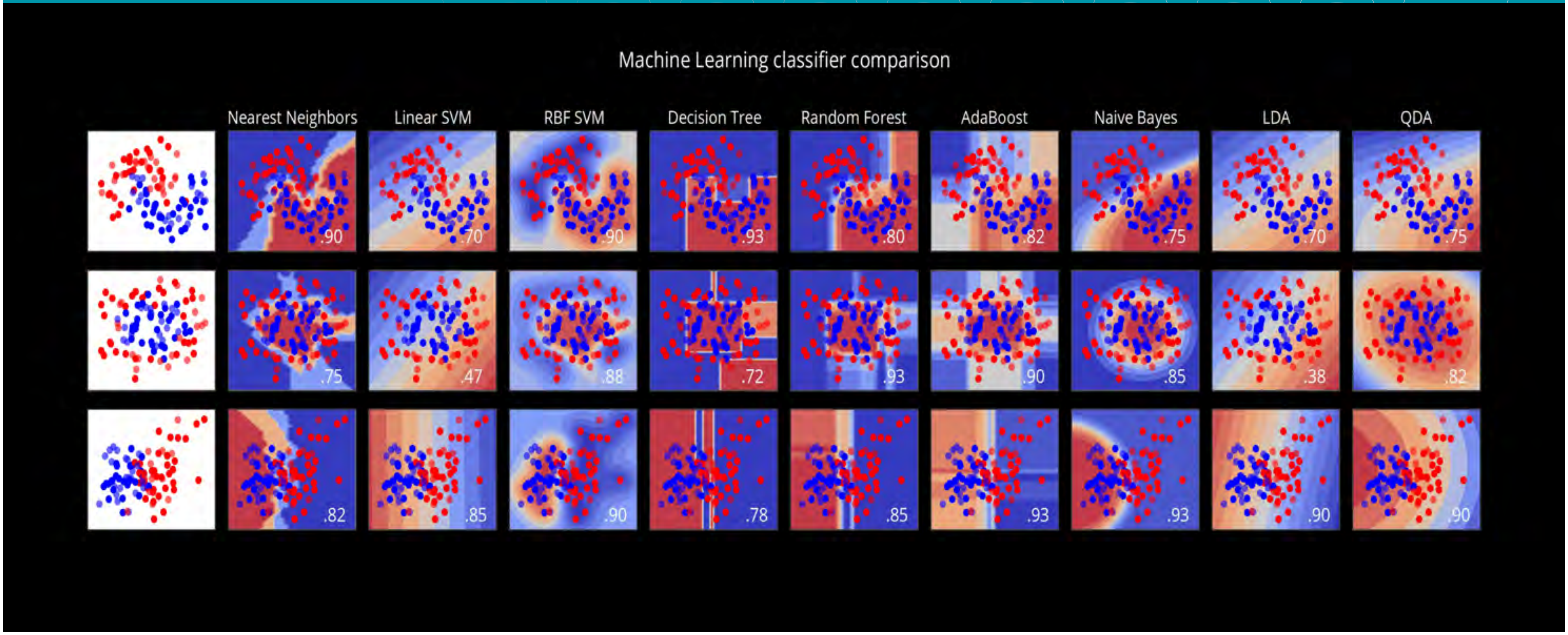
National Institute of Environmental Health Sciences
Your Environment. Your Health.

NIEHS Areas of Scientific Focus

Environmental Factor - July 2022: New initiatives to transform research highlighted at Council meeting (nih.gov)

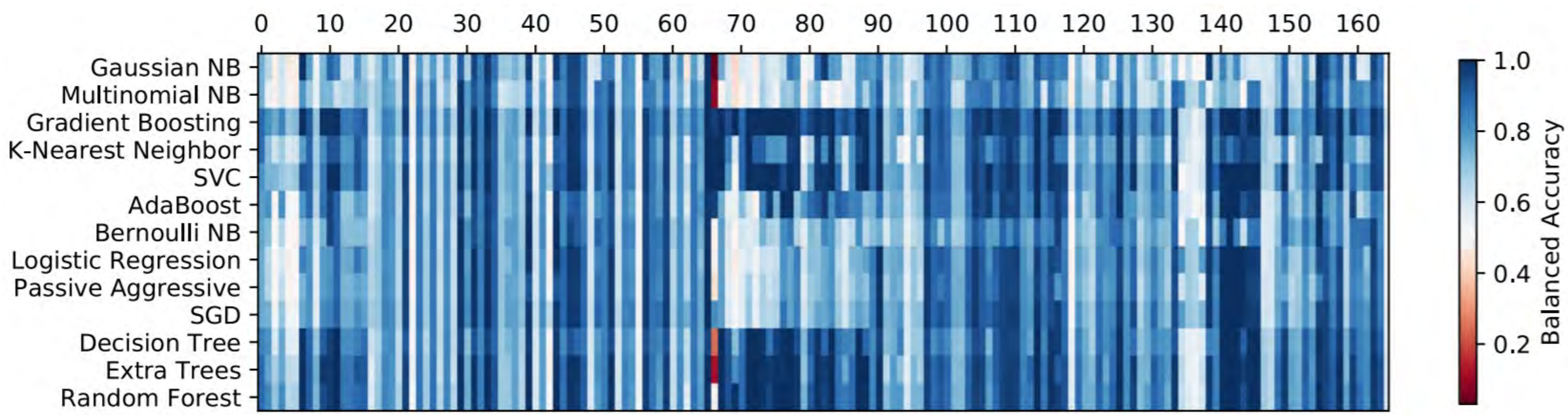

Easy Data Access
Easy Data Cleaning
Easy Data Analytics
Easy Model Evaluation
Easy Model Deployment
7-ELEVEN express
7-ELEVEN EXPRESS 새로운 경험
express

The practice of machine learning is hard!

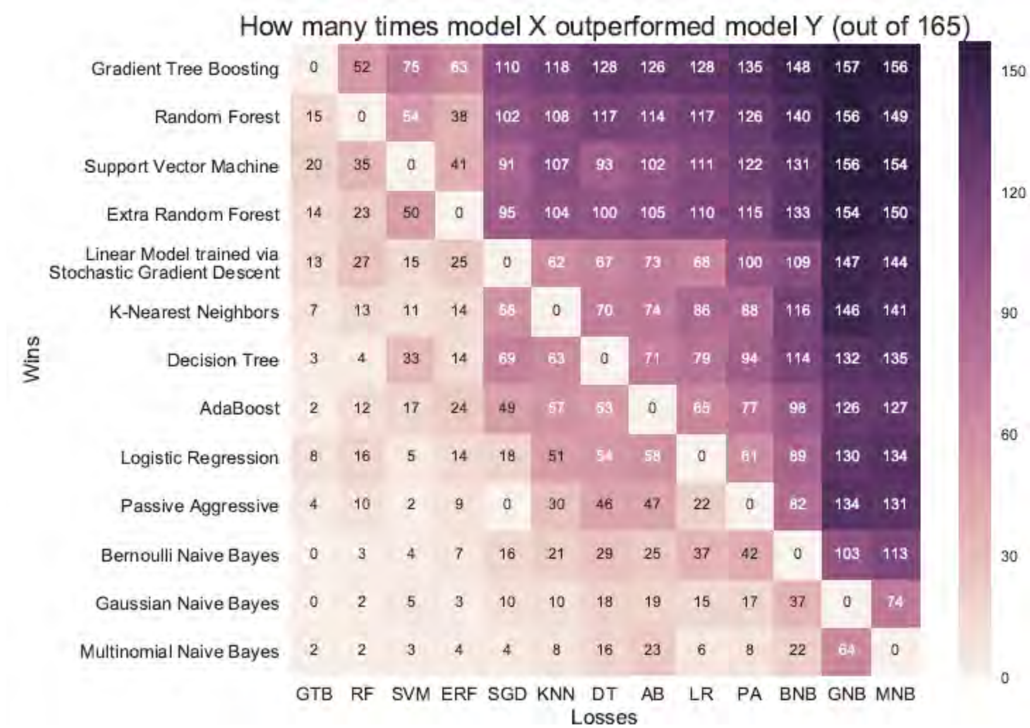


Made in Python by *Étienne Tétreault-Pinard*

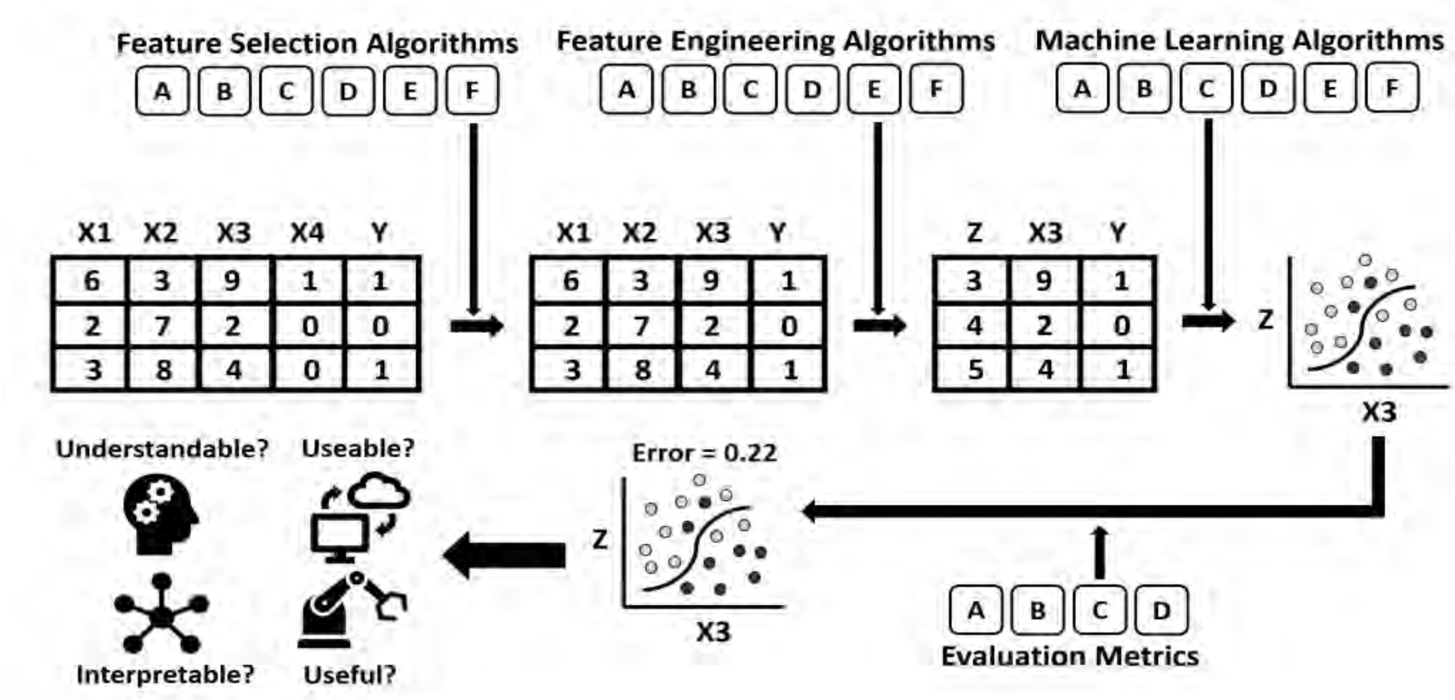
Machine learning is hard!



Machine learning is hard!



Building machine learning pipelines is harder



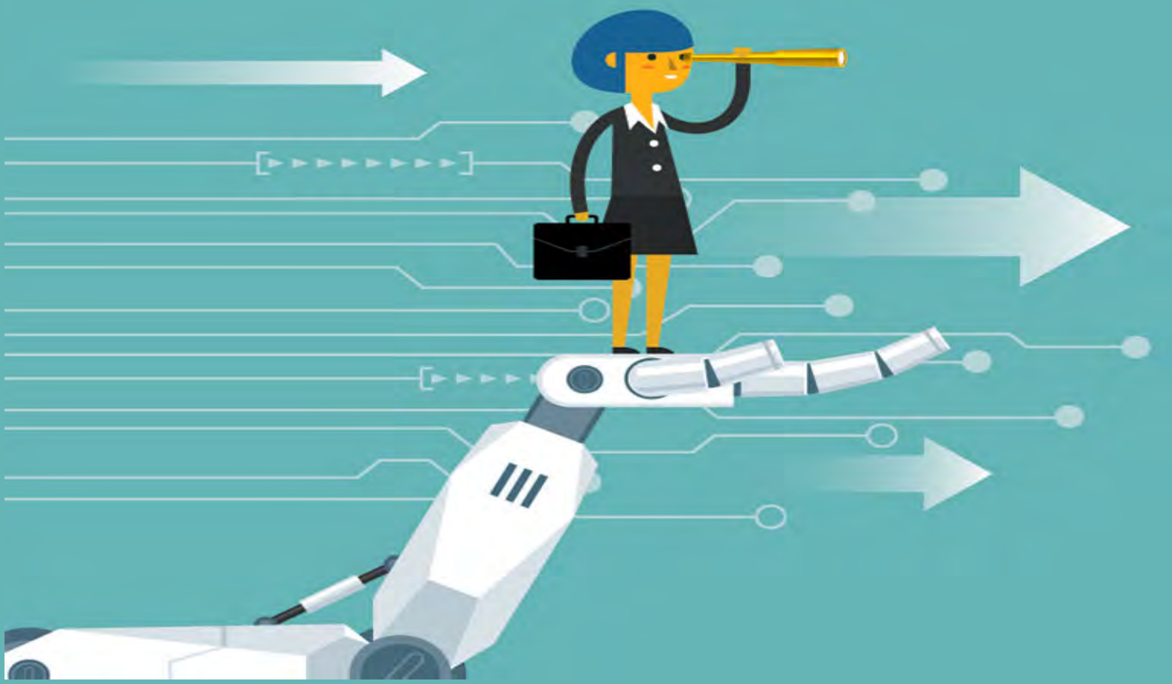
Deployment can be even harder

1. IT
2. UI/UX
3. Clinical workflow
4. Trust
5. Fairness
6. Ethics
7. ROI
8. Auditing
9. Legal
10. Updates

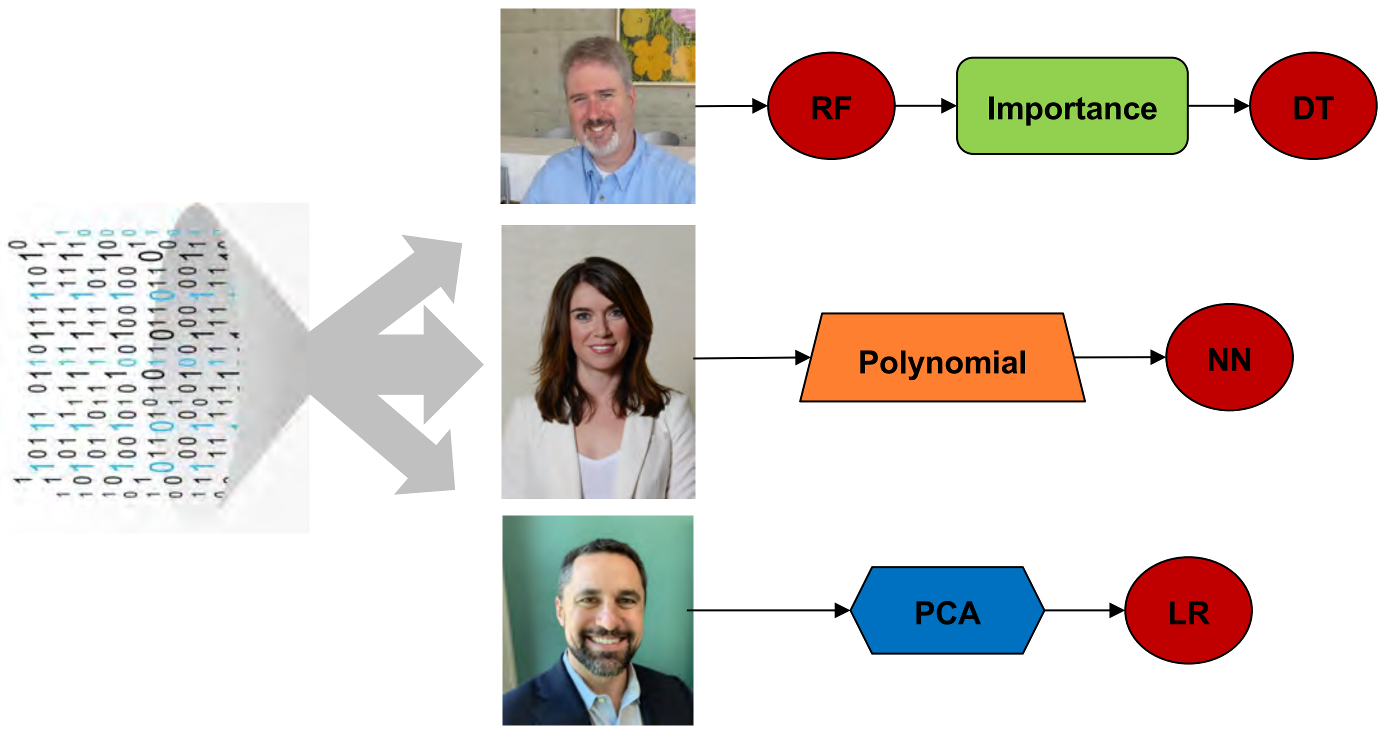
Cedars Sinai

GOAL: Democratize AI and machine learning so *everyone* can benefit

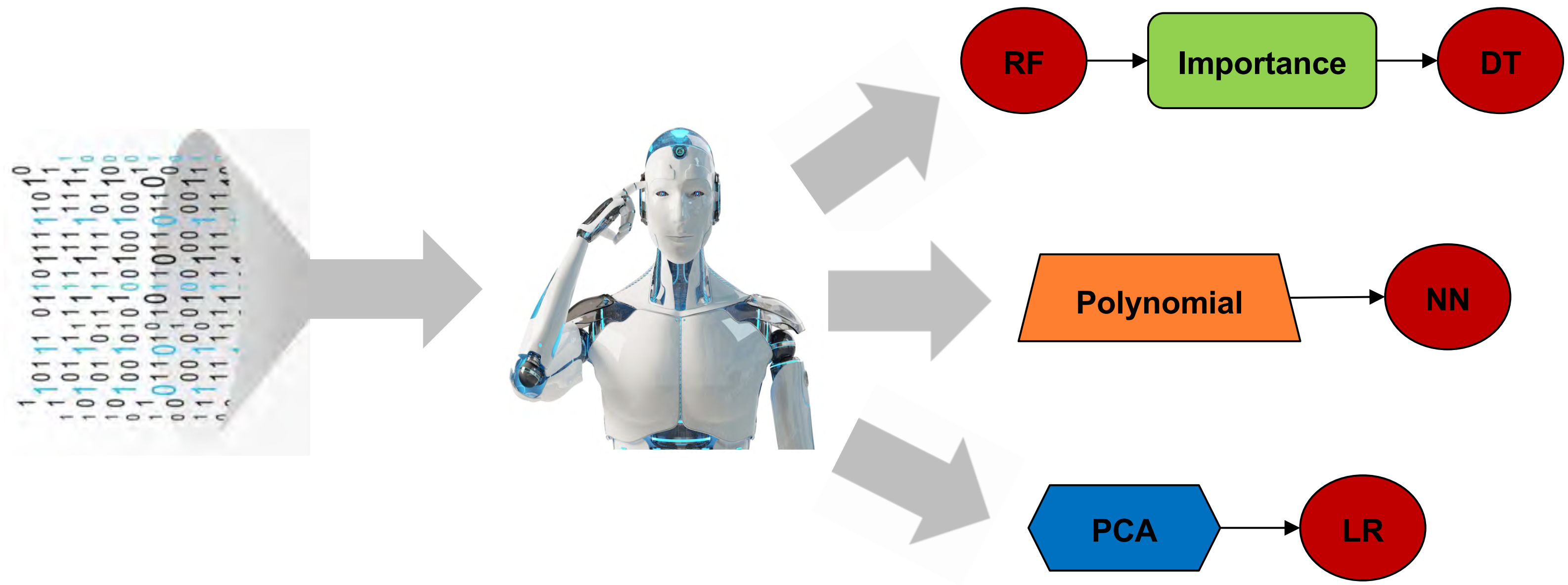
Necessary for a learning health system



How do humans build machine learning pipelines?



Can we get a computer to do this?



Tree-based pipeline optimization tool (TPOT)

TPOT

Automating data science

We were *two years* ahead of Google

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

Google is adding new automated machine learning tools and bringing its AI software to call centers

Google's AutoML Vision is going into public beta while new natural language and translation tools are now available

By Nick Statt | @nickstatt | Jul 24, 2018, 12:00pm EDT

The first book on automated machine learning (AutoML)

The Springer Series on Challenges in Machine Learning

Open Access

Automated Machine Learning

Methods, Systems, Challenges

Editors: **Hutter**, Frank, **Kotthoff**, Lars, **Vanschoren**, Joaquin (Eds.)

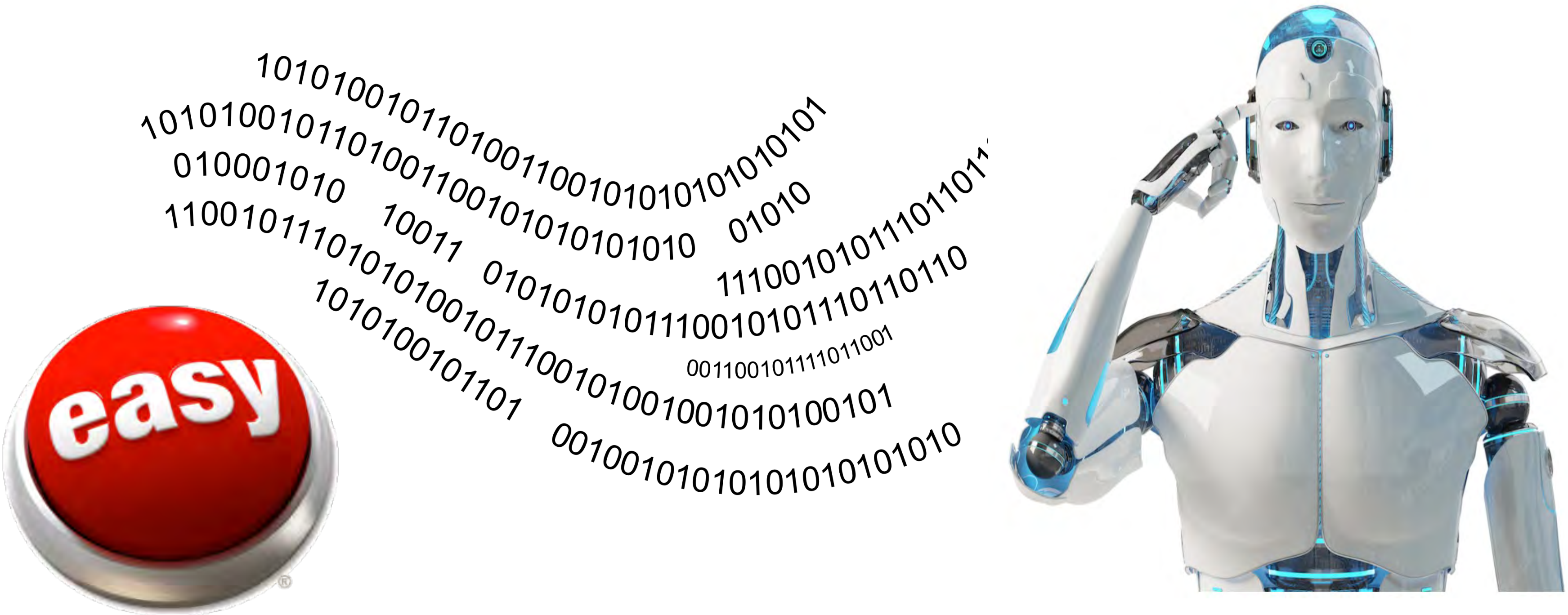
Presents a tutorial-level overview of the methods underlying automatic machine learning, enabling readers to easily understand the key concepts behind AutoML

Chapter 8
TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning

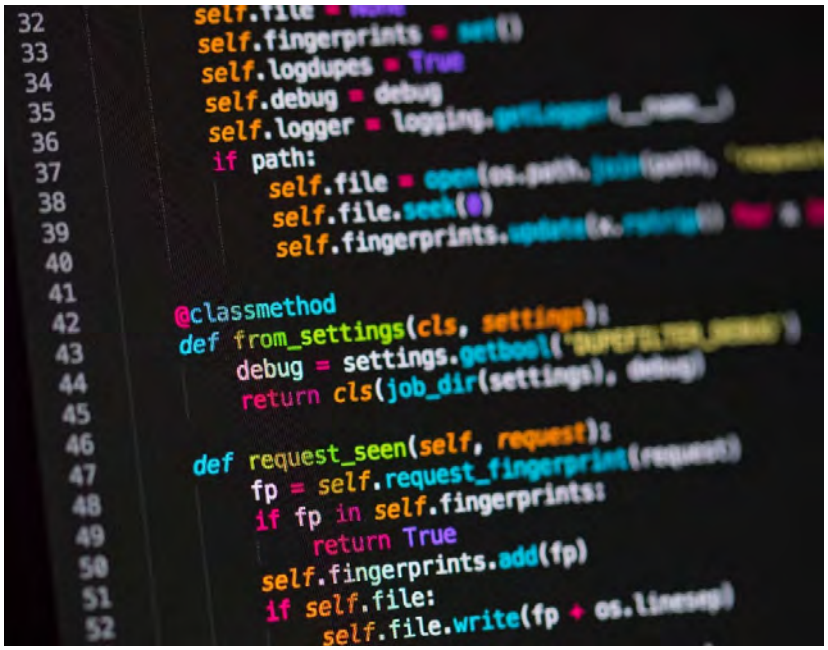
Randal S. Olson and Jason H. Moore

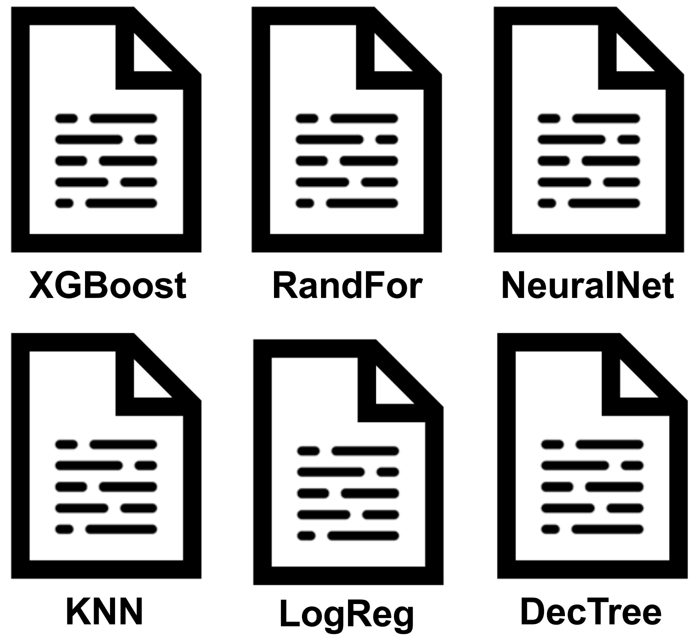
Abstract As data science becomes increasingly mainstream, there will be an ever-growing demand for data science tools that are more accessible, flexible, and scalable. In response to this demand, automated machine learning (AutoML) researchers have begun building systems that automate the process of designing and optimizing machine learning pipelines. In this chapter we present TPOT v0.3, an open source genetic programming-based AutoML system that optimizes a series of feature preprocessors and machine learning models with the goal of maximizing classification accuracy on a supervised classification task. We benchmark TPOT on a series of 150 supervised classification tasks and find that it significantly outperforms a basic machine learning analysis in 21 of them, while experiencing minimal degradation in accuracy on 4 of the benchmarks—all without any domain knowledge nor human input. As such, genetic programming-based AutoML systems show considerable promise in the AutoML domain.

Aliro AI for automated machine learning (AutoML)



Step 1 – Machine learning code base

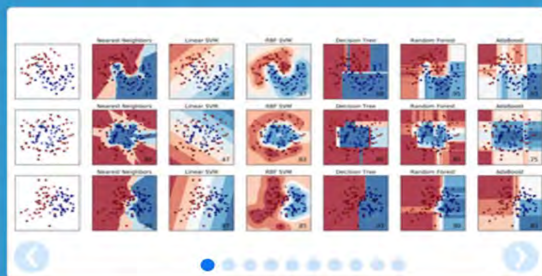




scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

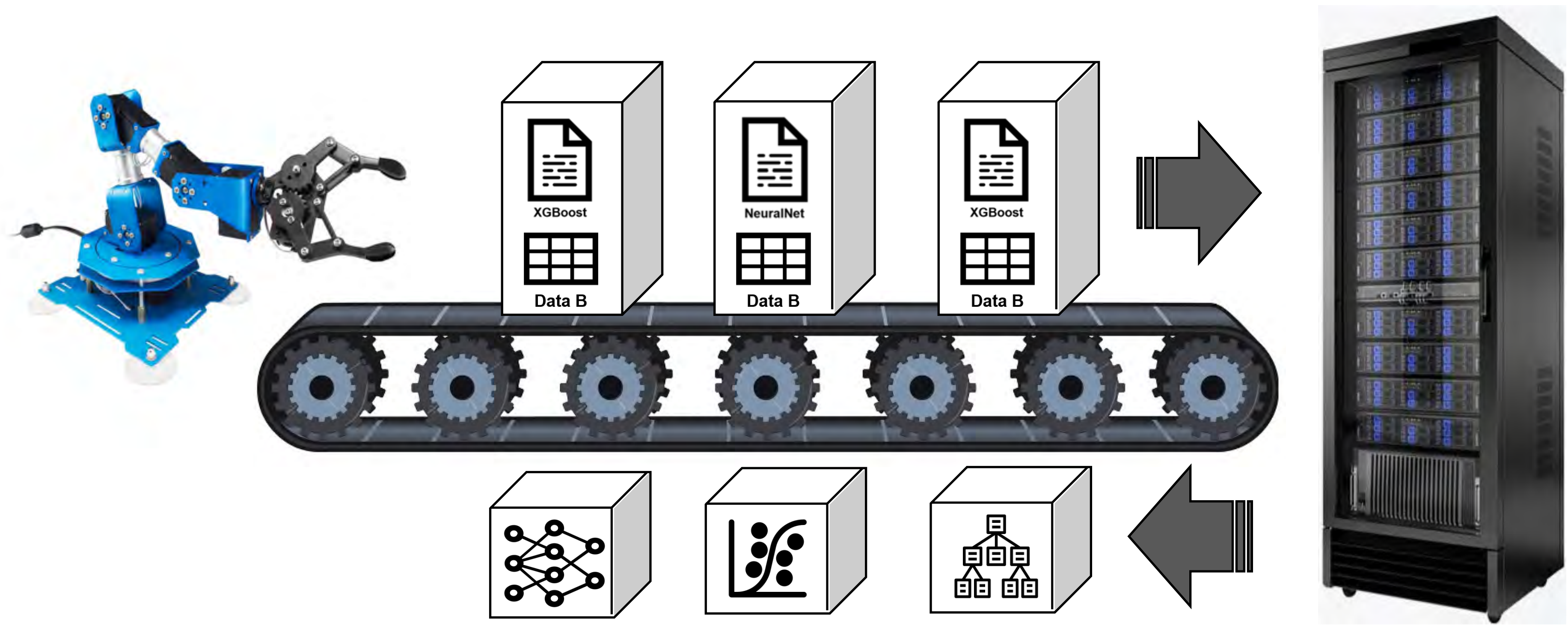
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

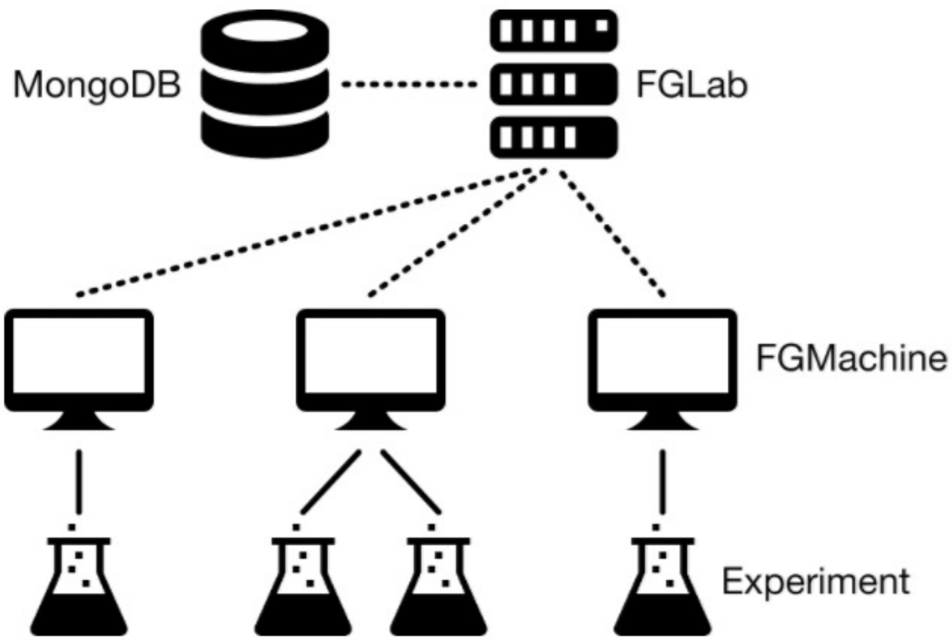
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

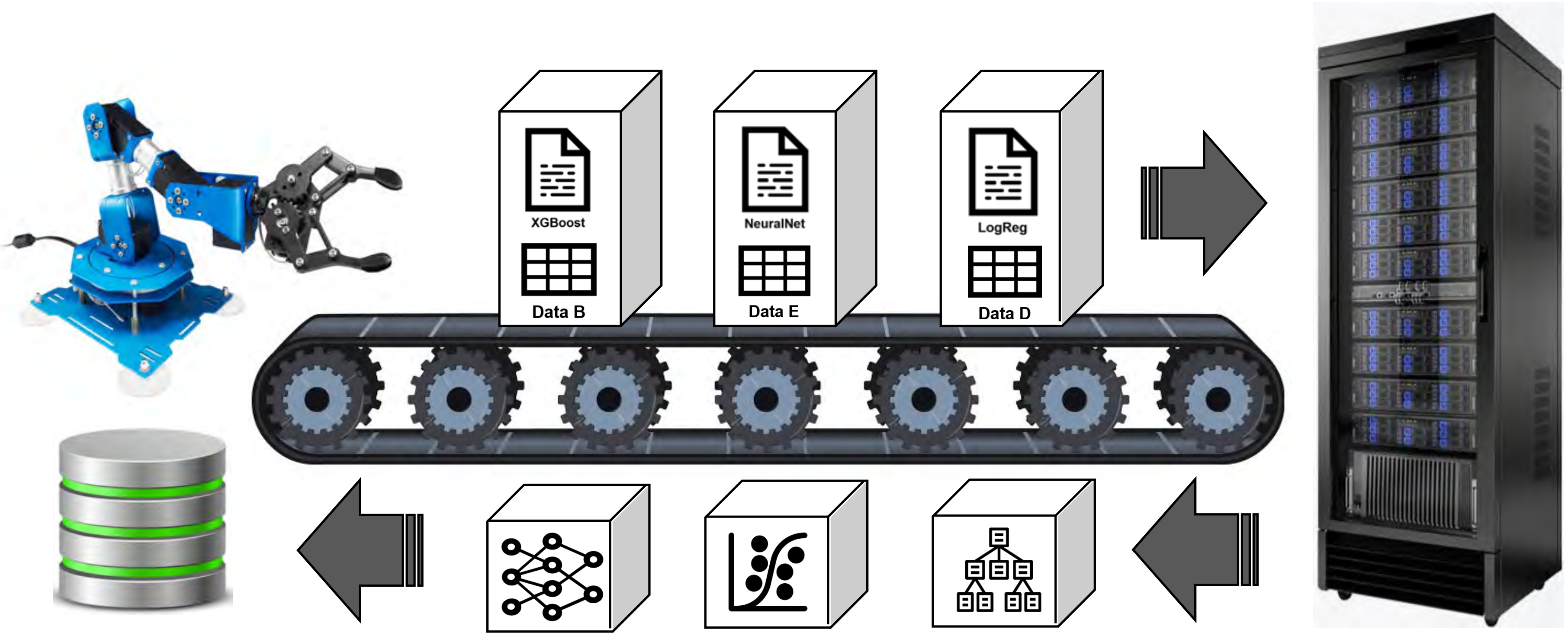
Step 2 – Controller



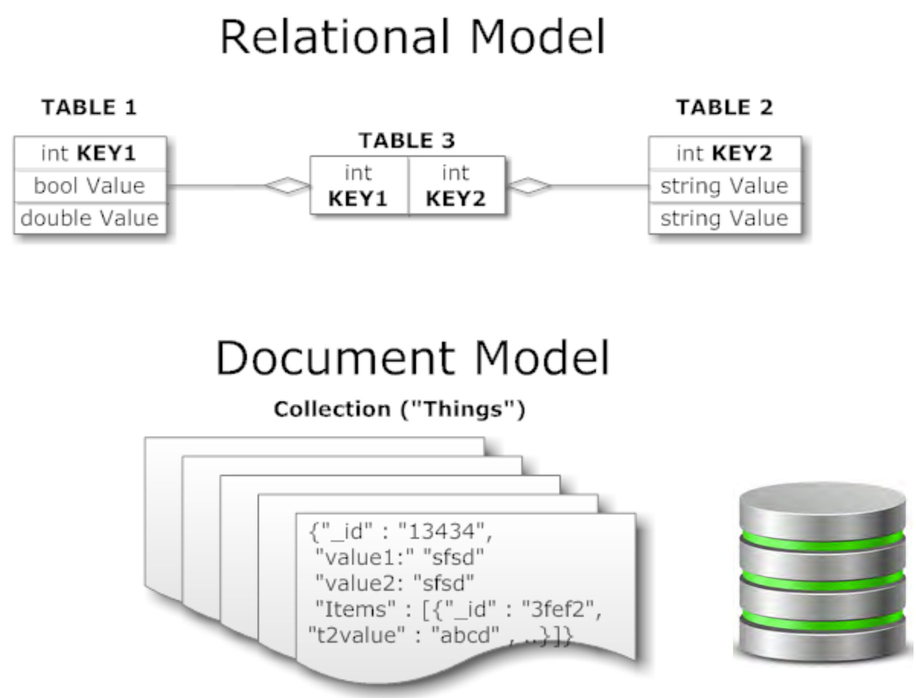
Future Gadget Lab



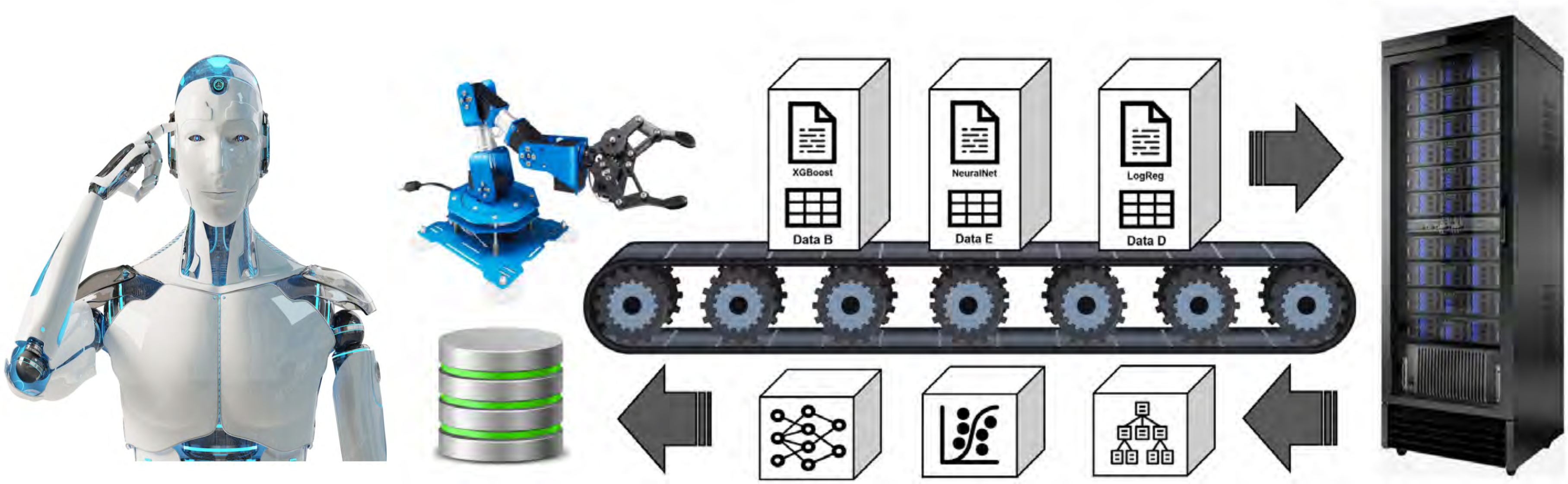
Step 3 – Database (i.e. the AI memory)



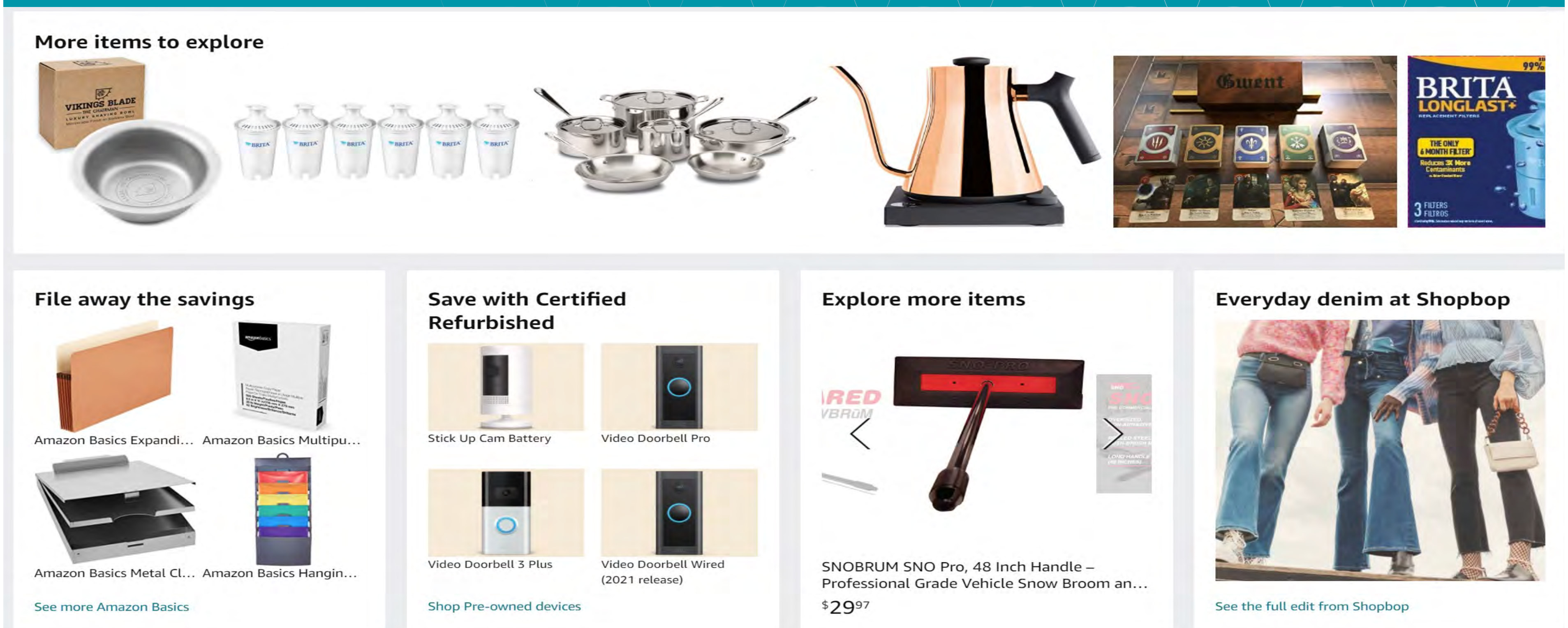
MongoDB Document Store



Step 4 – AI (i.e. something that learns from experience)



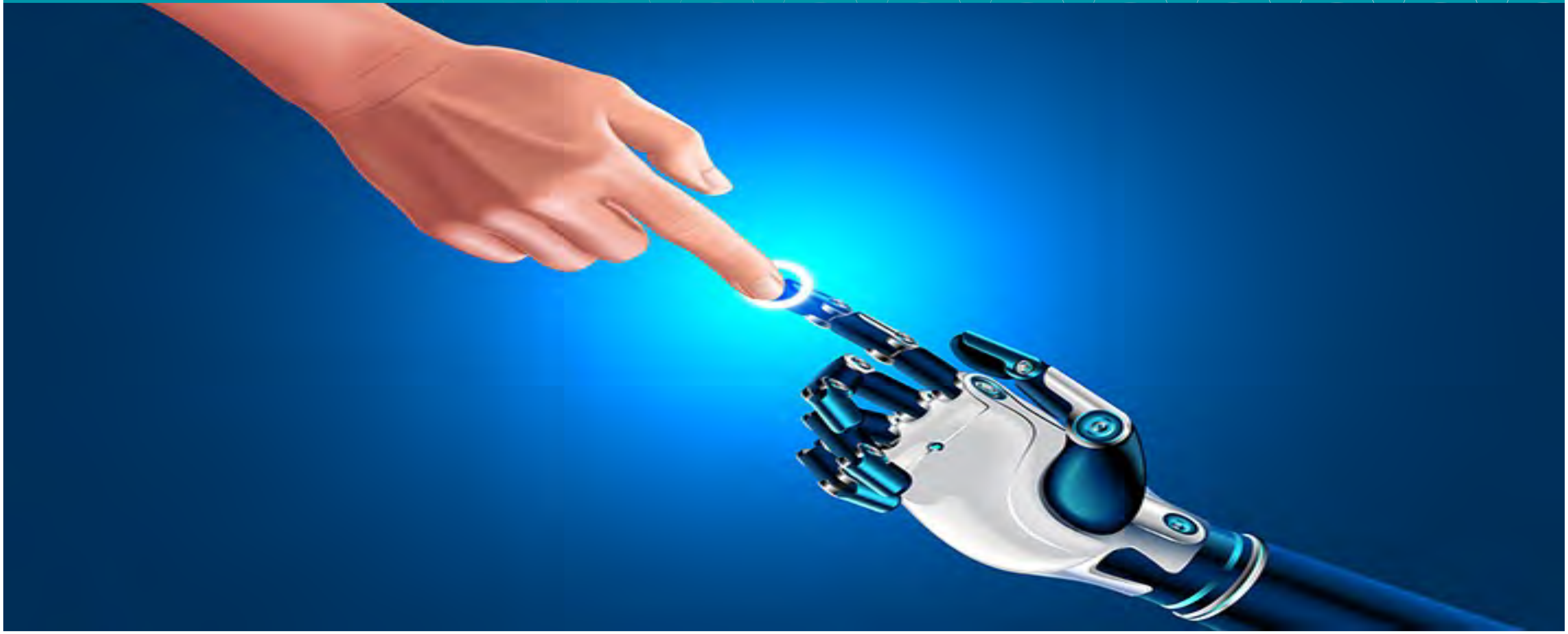
AI Recommender System

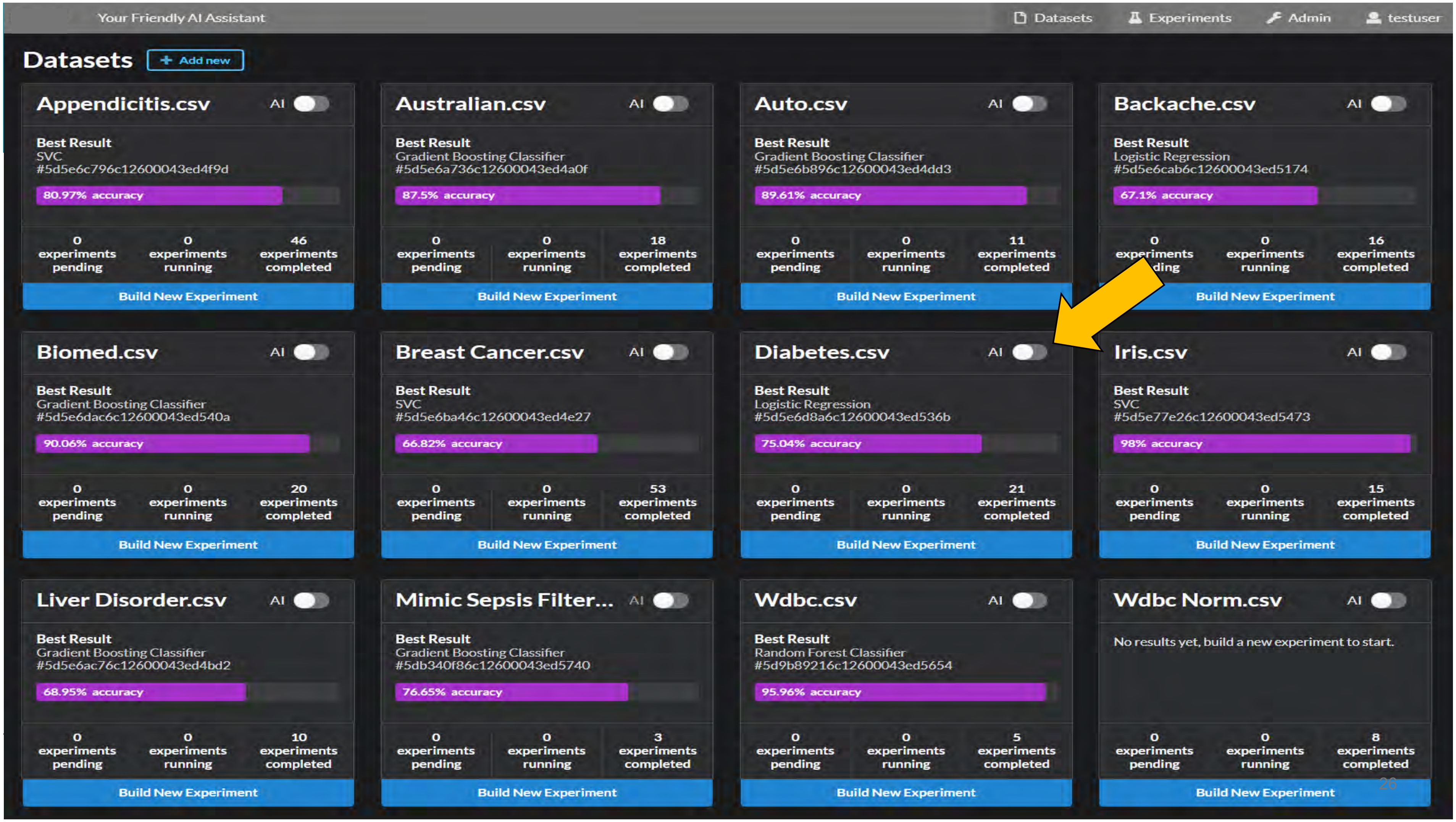


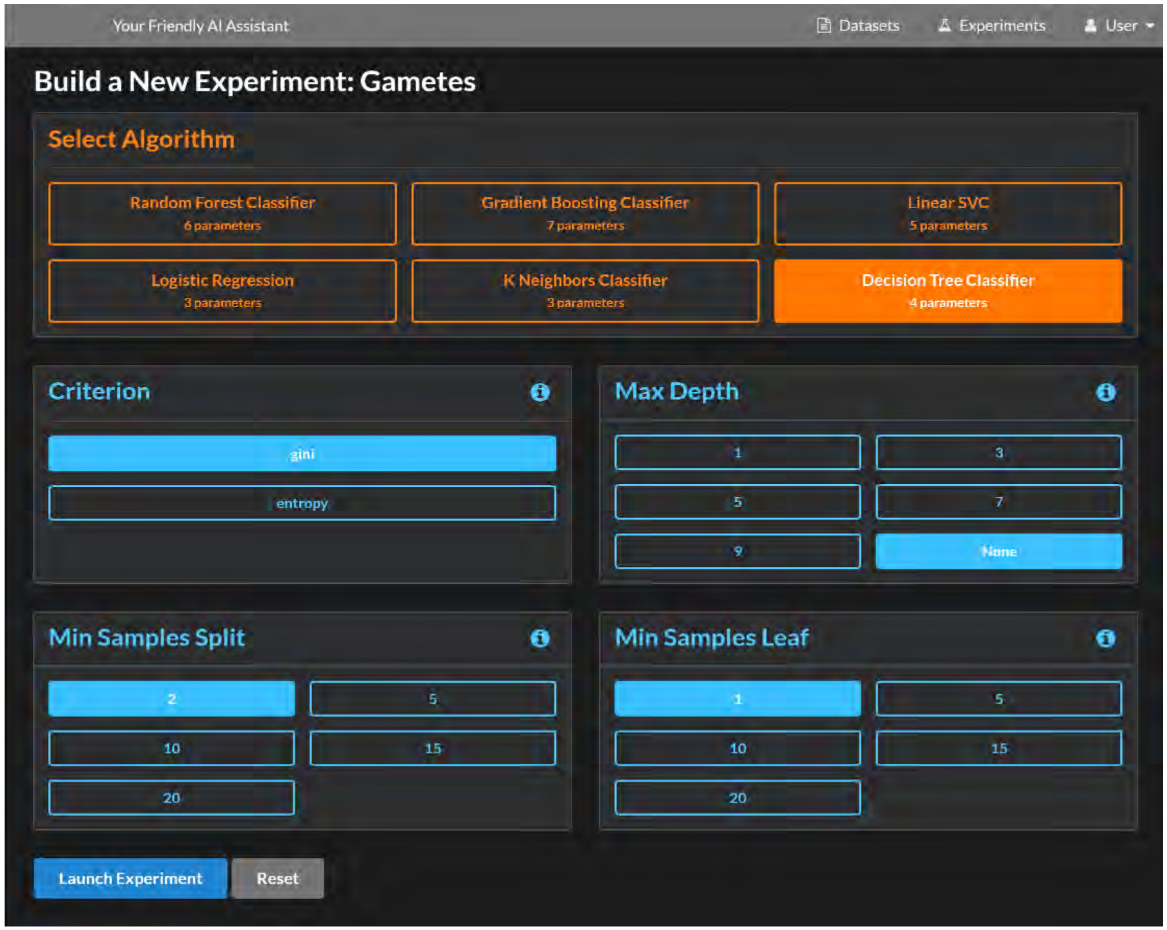
Recommender System

Result	Dataset	MF1	MF2	ML1P1	ML1P2	ML2P1	ML2P2	AUC
1	1	1	0	1	0	0	0	0.56
2	1	1	0	0	1	0	0	0.52
3	1	1	0	0	0	1	0	0.61
4	1	1	0	0	0	0	1	0.55
5	2	1	1	0	0	0	1	0.69
6	2	1	1	0	1	0	0	0.72
7	2	1	1	1	0	0	0	0.69
8	3	0	1	1	0	0	0	0.57
9	3	0	1	0	1	0	0	0.67
10	3	0	1	0	0	1	0	0.68
11	4	1	0	0	0	1	0	0.56
12	4	1	0	0	1	0	0	0.65
13	4	1	0	1	0	0	0	0.67
14	4	1	0	0	0	0	1	0.53
15	5	1	1	0	0	0	1	0.53
16	5	1	1	0	1	0	0	0.66
17	5	1	1	1	0	0	0	0.78
18	6	0	1	1	0	0	0	0.59
19	6	0	1	0	1	0	0	0.58
20	6	0	1	0	0	1	0	0.67

Step 5 – Human-Computer Interaction

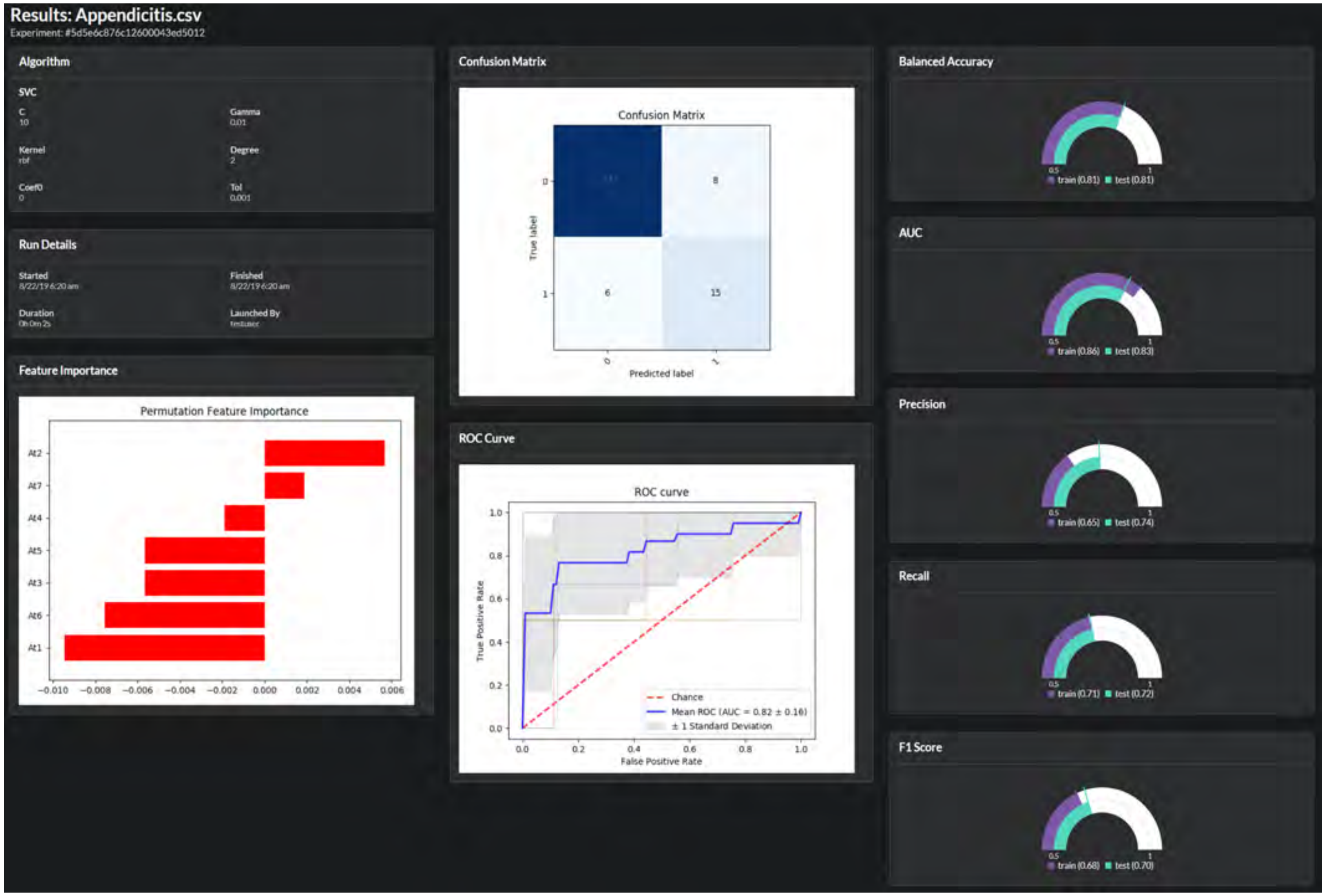



Your Friendly AI Assistant
Datasets
Experiments
Admin
testuser
Datasets
+ Add new
Appendicitis.csv
AI
Best Result
SVC
#5d5e6c796c12600043ed4f9d
80.97% accuracy
0 experiments pending
0 experiments running
46 experiments completed
Build New Experiment
Australian.csv
AI
Best Result
Gradient Boosting Classifier
#5d5e6a736c12600043ed4a0f
87.5% accuracy
0 experiments pending
0 experiments running
18 experiments completed
Build New Experiment
Auto.csv
AI
Best Result
Gradient Boosting Classifier
#5d5e6b896c12600043ed4dd3
89.61% accuracy
0 experiments pending
0 experiments running
11 experiments completed
Build New Experiment
Backache.csv
AI
Best Result
Logistic Regression
#5d5e6cab6c12600043ed5174
67.1% accuracy
0 experiments pending
0 experiments running
16 experiments completed
Build New Experiment
Biomed.csv
AI
Best Result
Gradient Boosting Classifier
#5d5e6dac6c12600043ed540a
90.06% accuracy
0 experiments pending
0 experiments running
20 experiments completed
Build New Experiment
Breast Cancer.csv
AI
Best Result
SVC
#5d5e6ba46c12600043ed4e27
66.82% accuracy
0 experiments pending
0 experiments running
53 experiments completed
Build New Experiment
Diabetes.csv
AI
Best Result
Logistic Regression
#5d5e6d8a6c12600043ed536b
75.04% accuracy
0 experiments pending
0 experiments running
21 experiments completed
Build New Experiment
Iris.csv
AI
Best Result
SVC
#5d5e77e26c12600043ed5473
98% accuracy
0 experiments pending
0 experiments running
15 experiments completed
Build New Experiment
Liver Disorder.csv
AI
Best Result
Gradient Boosting Classifier
#5d5e6ac76c12600043ed4bd2
68.95% accuracy
0 experiments pending
0 experiments running
10 experiments completed
Build New Experiment
Mimic Sepsis Filter...
AI
Best Result
Gradient Boosting Classifier
#5db340f86c12600043ed5740
76.65% accuracy
0 experiments pending
0 experiments running
3 experiments completed
Build New Experiment
Wdbc.csv
AI
Best Result
Random Forest Classifier
#5d9b89216c12600043ed5654
95.96% accuracy
0 experiments pending
0 experiments running
5 experiments completed
Build New Experiment
Wdbc Norm.csv
AI
No results yet, build a new experiment to start.
0 experiments pending
0 experiments running
8 experiments completed
Build New Experiment


Your Friendly AI Assistant
Datasets
Experiments
User
Build a New Experiment: Gametes
Select Algorithm
Random Forest Classifier
6 parameters
Gradient Boosting Classifier
7 parameters
Linear SVC
5 parameters
Logistic Regression
3 parameters
K Neighbors Classifier
3 parameters
Decision Tree Classifier
4 parameters
Criterion
gini
entropy
Max Depth
1
3
5
7
9
None
Min Samples Split
2
5
10
15
20
Min Samples Leaf
1
5
10
15
20
Launch Experiment
Reset

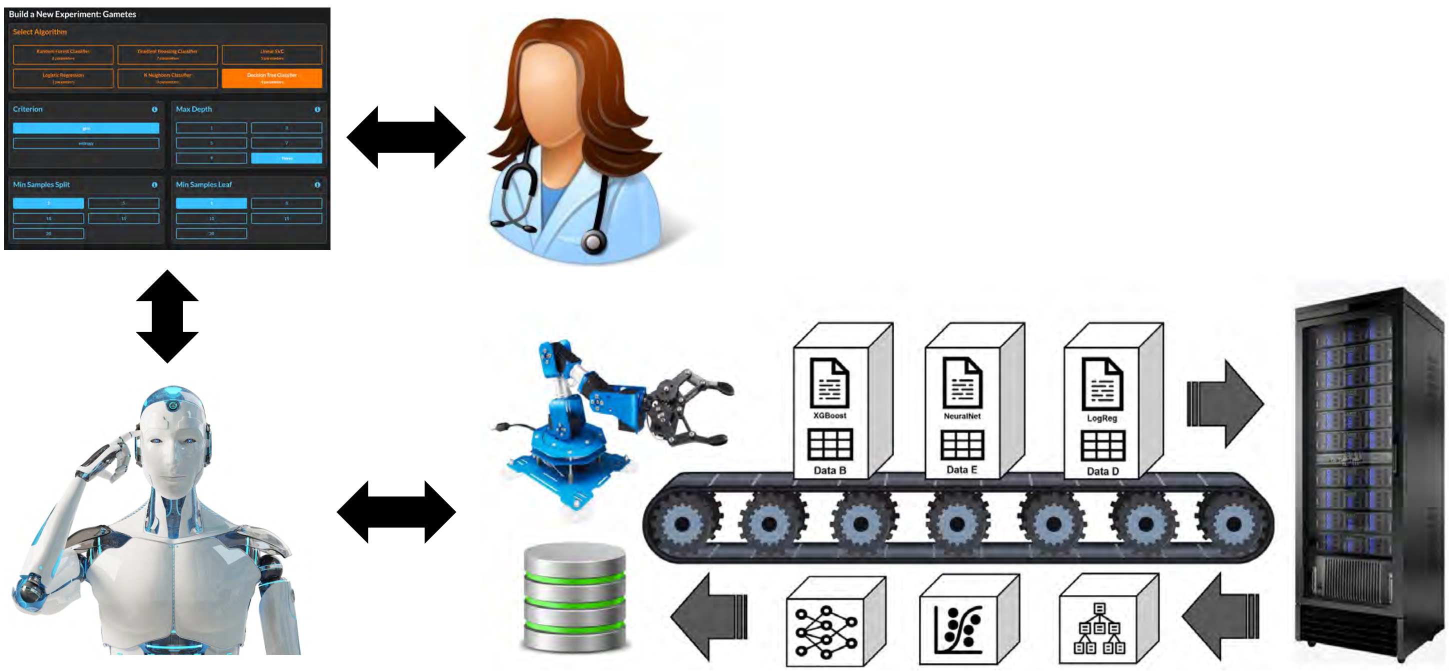
Status: Success Dataset: Adult Algorithm: All Reset filters **10 experiments**

Start Time	Accuracy	Dataset	Algorithm	Actions
12/10/18 11:36 am	0.83	Adult	Random Forest Classifier	
12/10/18 11:37 am	0.83	Adult	Random Forest Classifier	
12/10/18 12:27 pm	0.83	Adult	Random Forest Classifier	
12/10/18 11:27 am	0.83	Adult	Random Forest Classifier	
12/10/18 11:37 am	0.80	Adult	Gradient Boosting Classifier	
12/10/18 12:20 pm	0.79	Adult	Gradient Boosting Classifier	
12/10/18 12:29 pm	0.75	Adult	Gradient Boosting Classifier	
12/10/18 11:38 am	0.73	Adult	Gradient Boosting Classifier	
12/10/18 12:22 pm	0.72	Adult	Gradient Boosting Classifier	
12/10/18 12:20 pm	0.69	Adult	Gradient Boosting Classifier	

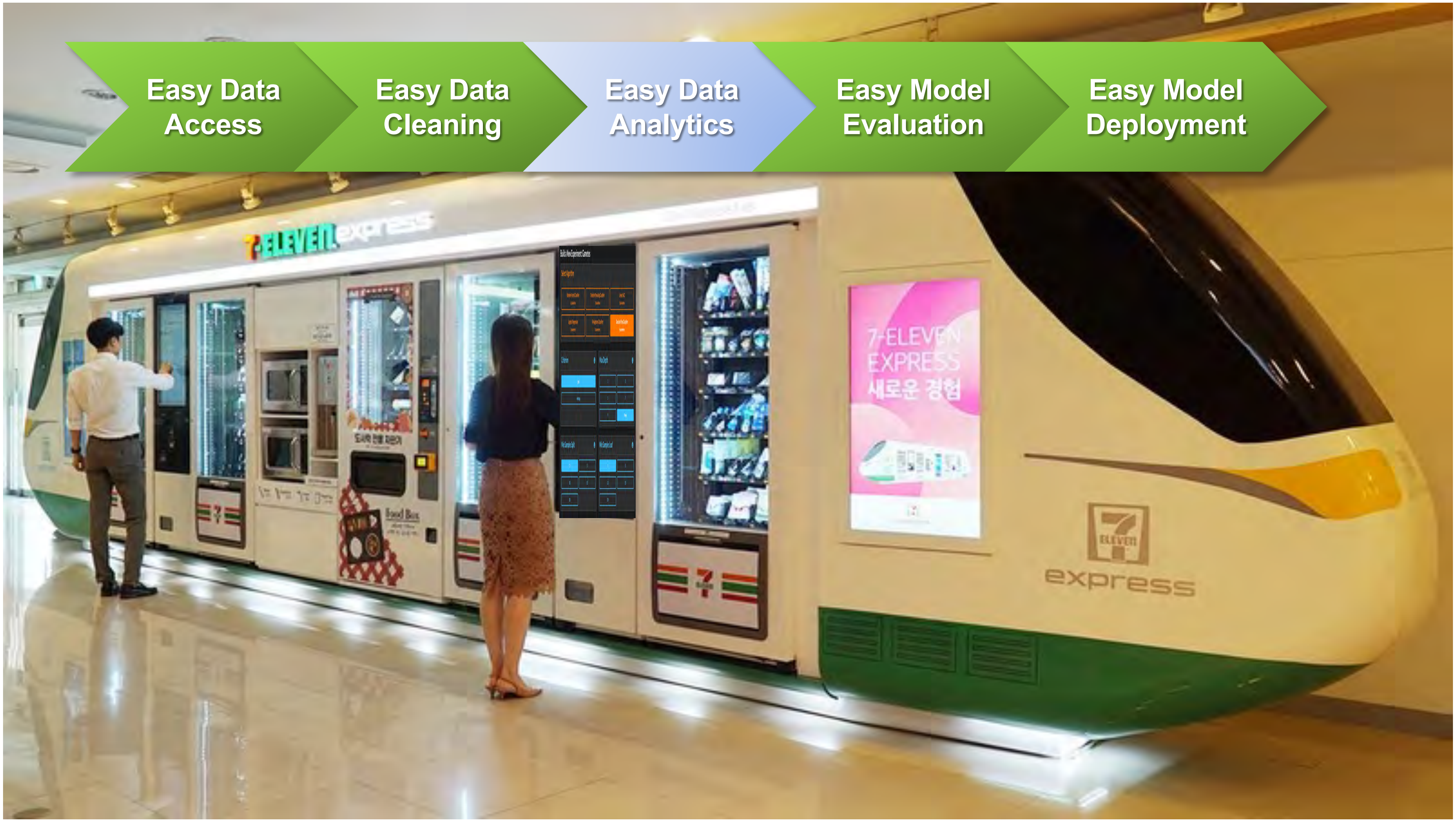


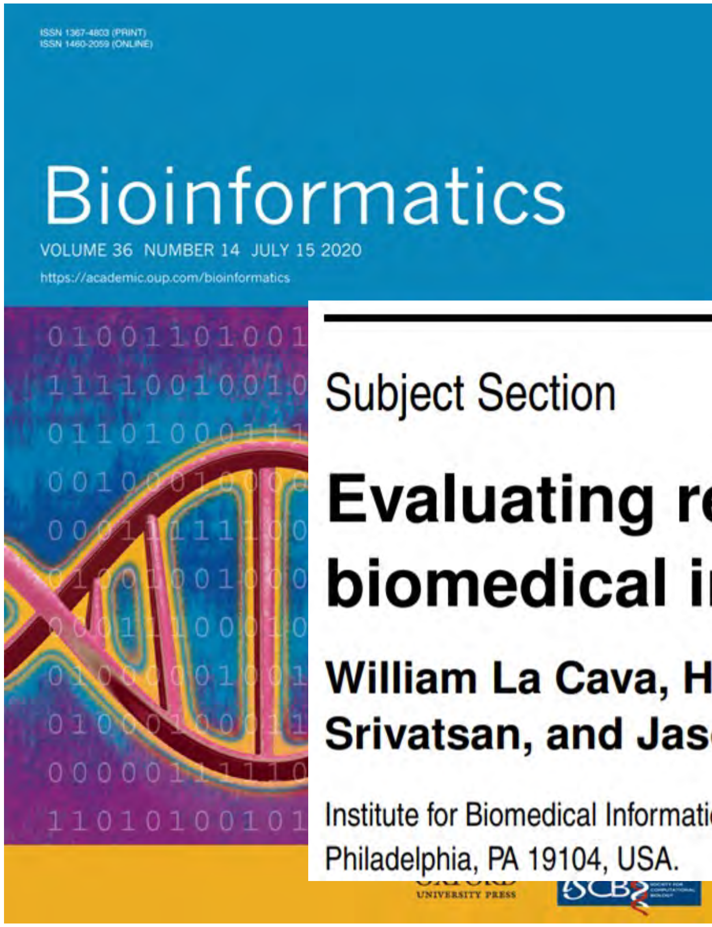
Results: Appendicitis.csv
Experiment: #5d5e6c876c12600043ed5012
Algorithm
SVC
C
10
Gamma
0.01
Kernel
rbf
Degree
2
Coef0
0
Tol
0.001
Run Details
Started
8/22/19 6:20 am
Finished
8/22/19 6:20 am
Duration
0h 0m 2s
Launched By
Feature Importance
Permutation Feature Importance
At2
At7
At4
At5
At3
At6
At1
Confusion Matrix
Confusion Matrix
True label
Predicted label
8
6
15
ROC Curve
ROC curve
True Positive Rate
False Positive Rate
Chance
Mean ROC (AUC = 0.82 ± 0.16)
± 1 Standard Deviation
Balanced Accuracy
train (0.81)
test (0.81)
AUC
train (0.86)
test (0.83)
Precision
train (0.65)
test (0.74)
Recall
train (0.71)
test (0.72)
F1 Score
train (0.68)
test (0.70)

Aliro AI



Build a New Experiment: Gametes
Select Algorithm
Criterion
Max Depth
Min Samples Split
Min Samples Leaf
XGBoost
Data B
NeuralNet
Data E
LogReg
Data D


Easy Data Access
Easy Data Cleaning
Easy Data Analytics
Easy Model Evaluation
Easy Model Deployment
7-ELEVEn express
7-ELEVEN EXPRESS
새로운 경험
Food Box
7 ELEVEn
express

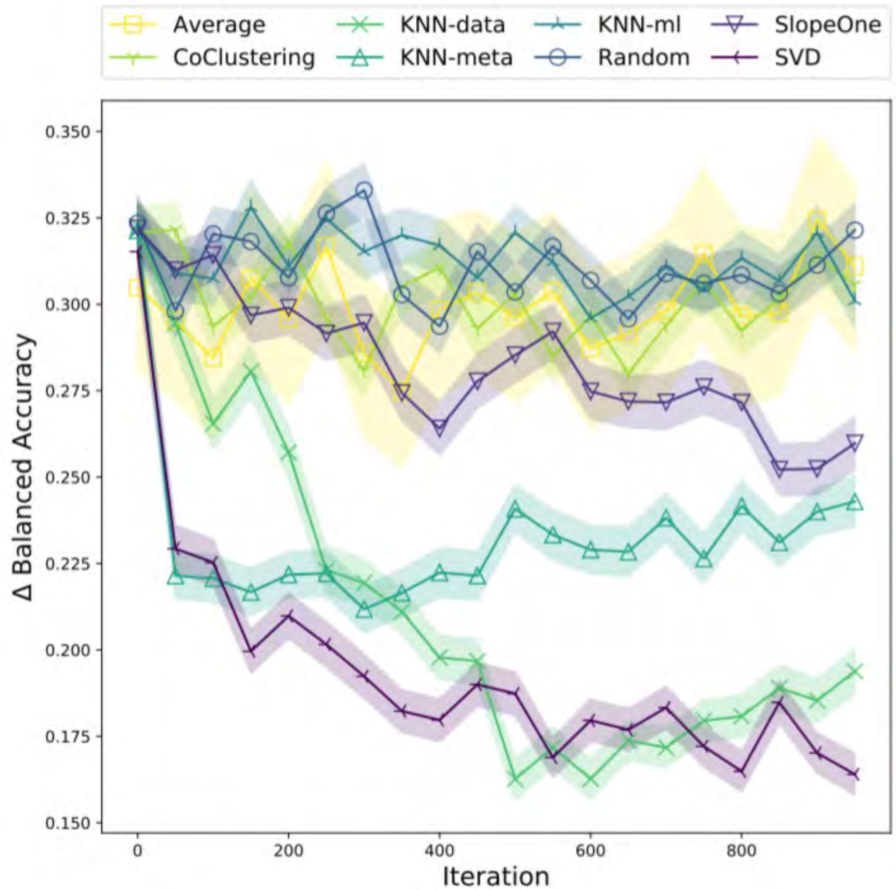


Subject Section

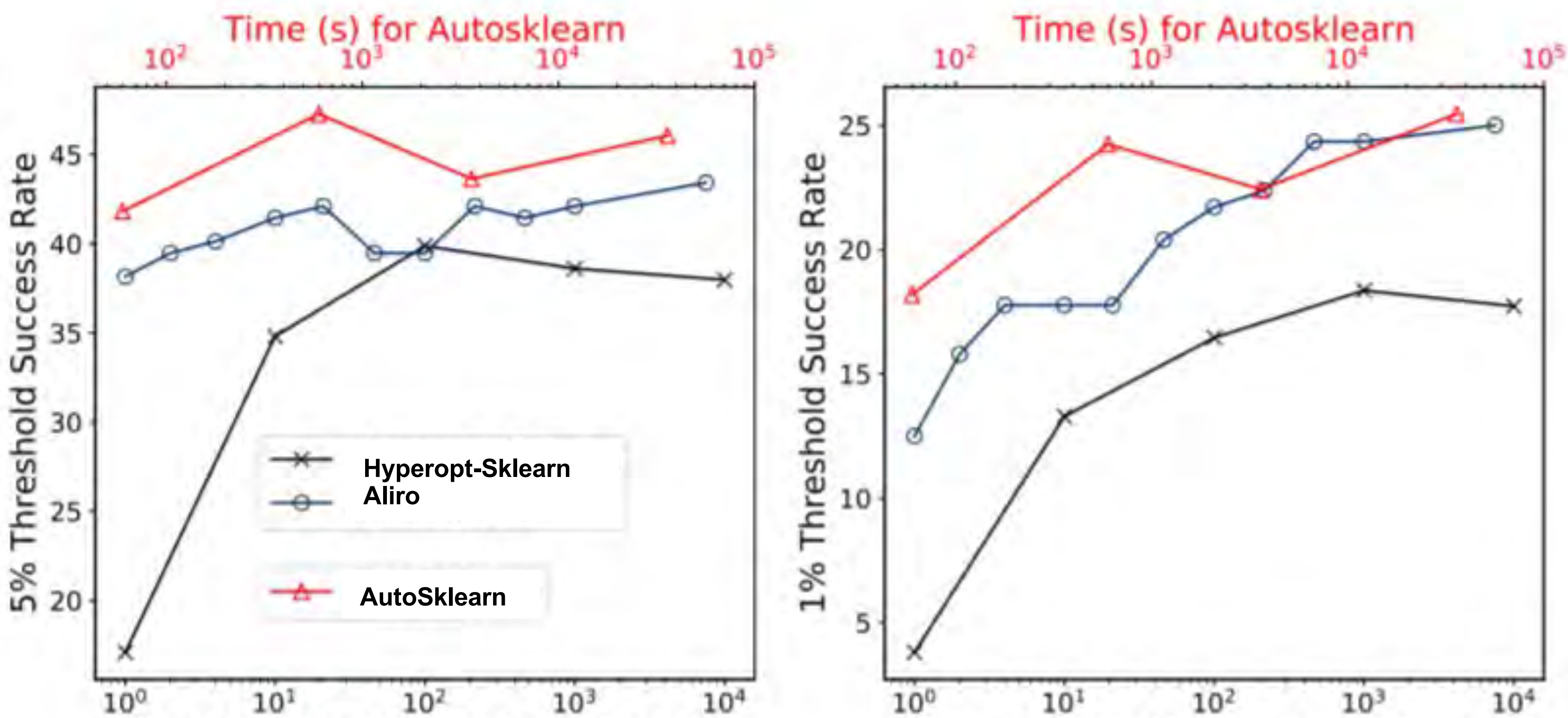
Evaluating recommender systems for AI-driven biomedical informatics

William La Cava, Heather Williams, Weixuan Fu, Steve Vitale, Durga Srivatsan, and Jason H. Moore*

Institute for Biomedical Informatics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA.



Average
CoClustering
KNN-data
KNN-meta
KNN-ml
Random
SlopeOne
SVD
Δ Balanced Accuracy
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0
200
400
600
800
Iteration


Time (s) for Autosklearn
5% Threshold Success Rate
Hyperopt-Sklearn
Aliro
AutoSklearn
Time (s) for Autosklearn
1% Threshold Success Rate

scientific reports

Explore our content ⌄ Journal information ⌄

nature › scientific reports › articles › article

Article | Open Access | Published: 22 October 2019

LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock

Josef Fagerström, Magnus Bång, Daniel Wilhelms & Michelle S. Chew

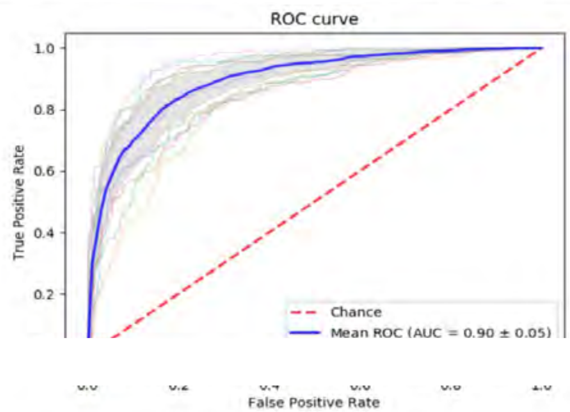
Scientific Reports **9**, Article number: 15132 (2019) | Cite this article

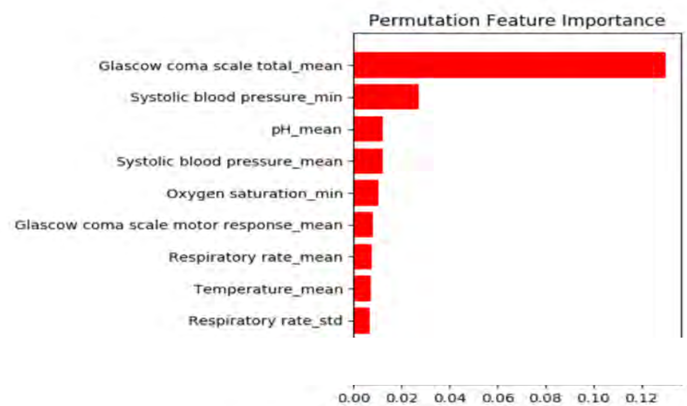
Abstract

Sepsis is a major health concern with global estimates of 31.5 million cases per year. Case fatality rates are still unacceptably high, and early detection and treatment is vital since it significantly reduces mortality rates for this condition. Appropriately designed automated detection tools have the potential to reduce the morbidity and mortality of sepsis by providing early and accurate identification of patients who are at risk of developing sepsis. In this paper, we present "LiSep LSTM"; a Long Short-Term Memory neural network designed for early identification of septic shock. LSTM networks are typically well-suited for detecting long-term dependencies in time series data. LiSep LSTM was developed using the machine learning framework Keras with a Google TensorFlow back end. The model was trained with data from the Medical Information Mart for Intensive Care database which contains vital signs, laboratory data, and journal entries from approximately 59,000 ICU patients. We show that LiSep LSTM can outperform a less complex model, using the same features and targets, with an AUROC 0.8306 (95% confidence interval: 0.8236, 0.8376) and median offsets between prediction and septic shock onset up to 40

learning algorithms for early detection of sepsis.

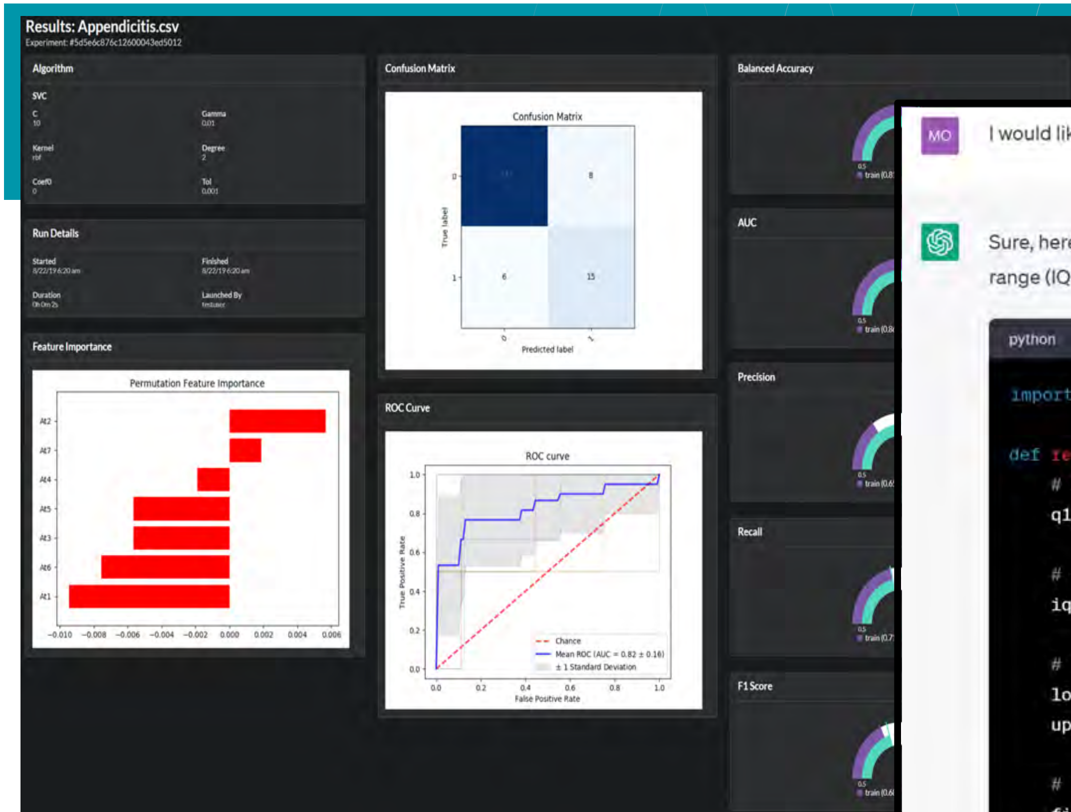
Model	Hyperparameters (Sklearn syntax)	Source	5-fold CV AUROC	Test AUROC	Training time
Logistic Regression	penalty=L2, C=1.0, Dual=False	user	0.84 ± 0.03	0.74 ± 0.03	13 s
	penalty=L1, C=1e-3, Dual=False	AI	0.81 ± 0.03	0.80 ± 0.02	16 s
KNN Classifier	n_neighbors=1, weights='distance'	AI	0.68 ± 0.02	0.67 ± 0.03	1 m 1 s
	n_neighbors=7, weights='uniform'	AI	0.81 ± 0.03	0.80 ± 0.03	11 m 3 s
	n_neighbors=11, weights='uniform'	AI	0.82 ± 0.03	0.81 ± 0.02	11 m 13 s
Support Vector	kernel=rbf, C=1.0, degree=3, gamma=0.01	user	0.85 ± 0.02	0.84 ± 0.02	7 m 55 s
	kernel=polynomial, C=1.0, degree=3, gamma=0.01	user	0.82 ± 0.01	0.83 ± 0.02	8 m 43 s
Classification Tree	criterion=entropy, max_depth=10	AI	0.75 ± 0.02	0.75 ± 0.03	5 s
	criterion=gini, max_depth=10	user	0.73 ± 0.06	0.59 ± 0.04	6 s
Random Forest	criterion=Gini, max_features='sqrt', n_estimators=100	user	0.88 ± 0.06	0.82 ± 0.02	2 m 7 s
	criterion=entropy, max_features='log2', n_estimators=100	AI	0.83 ± 0.06	0.74 ± 0.03	1 m 16 s
Gradient Boosting	**learning_rate=0.01, max_depth=5, n_estimators=500**	**AI**	**0.90 ± 0.05**	**0.85 ±0.02**	**1 m 58 s**
	learning_rate=0.1, max_depth=5, n_estimators=500	AI	0.89 ± 0.05	0.83 ± 0.02	2 m 25 s
	learning_rate=0.1, max_depth=3, n_estimators=500	AI	0.89 ± 0.05	0.84 ± 0.02	1 m 18 s
	learning_rate=1, max_depth=1, n_estimators=100	AI	0.88 ± 0.06	0.82 ± 0.03	10 s
Tuned Logistic Regression	penalty=L2, 10 C values log-spaced in [1e − 4, ..., 1e4]	control	0.84 ± 0.02	0.81 ± 0.02	1 m 8 s
LSTM	*dim=256, depth=1, dropout=0.3, 100 epochs*	*Harutyunyan* et al. *(2019)*	-	*0.86 ± 0.02*	*30 h 21 m*

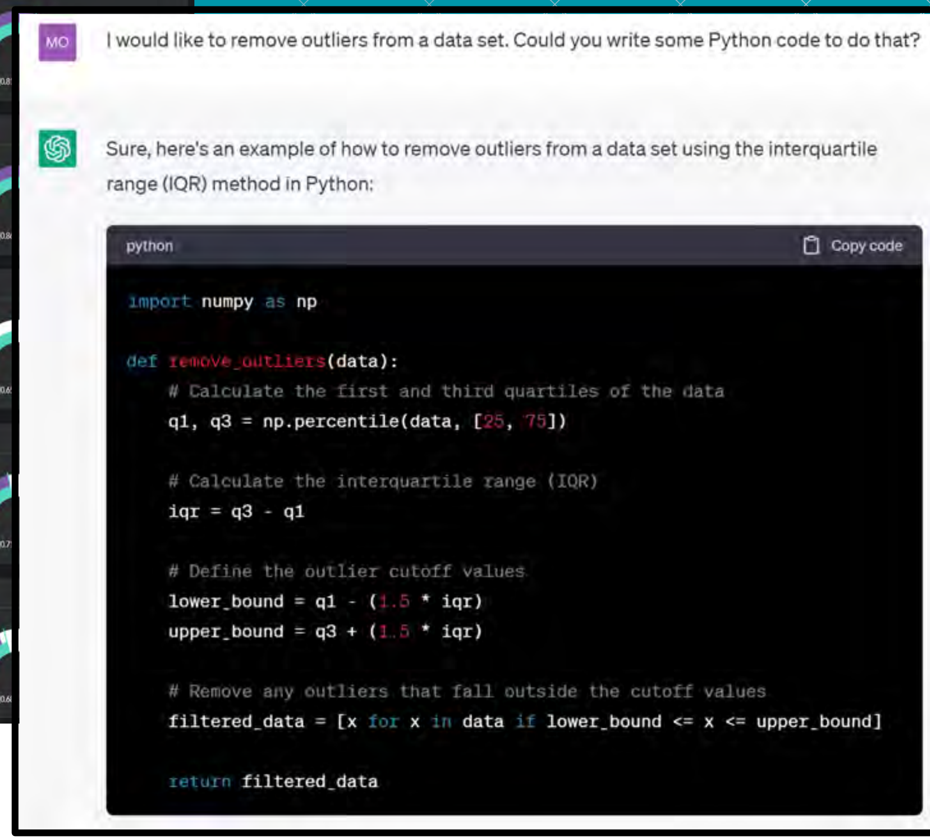




Aliro AI is open source

https://github.com/EpistasisLab/aliro





Aliro + ChatGPT

Question – How to integrate expert knowledge in an AI?

15 April 1983, Volume 220, Number 4594 SCIENCE

Expert Systems Research

Richard O. Duda and Edward H. Shortliffe

Few areas of research have been as exciting, promising, or bewildering as artificial intelligence (AI). After 25 years of use, the very name—combining as it does a highly immodest ambition with a cause they address problems normally thought to require human specialists for their solution (*2*). Some of these programs have reached expert levels of performance on the problems for which they that digital computers are not just fast adding machines, but are general-purpose processors of symbols, potentially capable of being programmed to exhibit such intelligent behavior (*3*). To support this view, AI researchers wrote programs to solve well-defined problems that had a distinctly nonnumerical character—programs that could play games, solve puzzles, perform symbolic integration, and even prove simple theorems in algebra, geometry, and symbolic logic. Among the important techniques that emerged were general methods for representing information in symbolic data structures, general methods for manipu-

Knowledge acquisition. One common characteristic of MYCIN, INTERNIST, and PROSPECTOR is that their knowledge bases are incomplete. The identification and encoding of knowledge is one of the most complex and arduous tasks encountered in the construction of an expert system. The very attempt to build a knowledge base often discloses gaps in our understanding of the subject domain and weaknesses in available representation techniques. Even when an adequate knowledge representation formalism has been developed, experts often have difficulty expressing their knowledge in that form. Thus the process of building a

Knowledge representation. MYCIN's knowledge about bacterial infections and R1's knowledge about computer configuration are represented by rules. The advantages and limitations of such simple, uniform approach are well appreciated by AI researchers, who have developed a variety of alternative formalisms for knowledge representation (*24*). The methods currently in use in expert systems rarely capture subtleties and sometimes fail to reflect major aspects of an expert's knowledge. For example, while MYCIN and INTERNIST-1 have effective mechanisms for representing empirical associations, neither has appropriate ways to express physiological mechanisms or temporal trends in the evolution of disease processes.

Computational Toxicology Knowledgebase (Comptox.AI)

Chemical Research in Toxicology

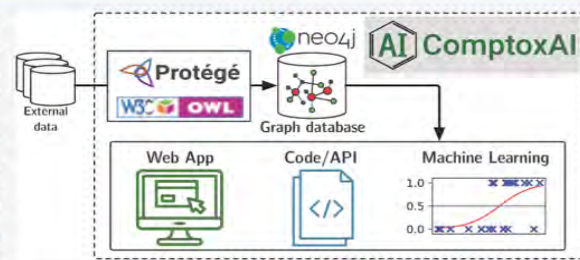
pubs.acs.org/crt Article

Automating Predictive Toxicology Using ComptoxAI

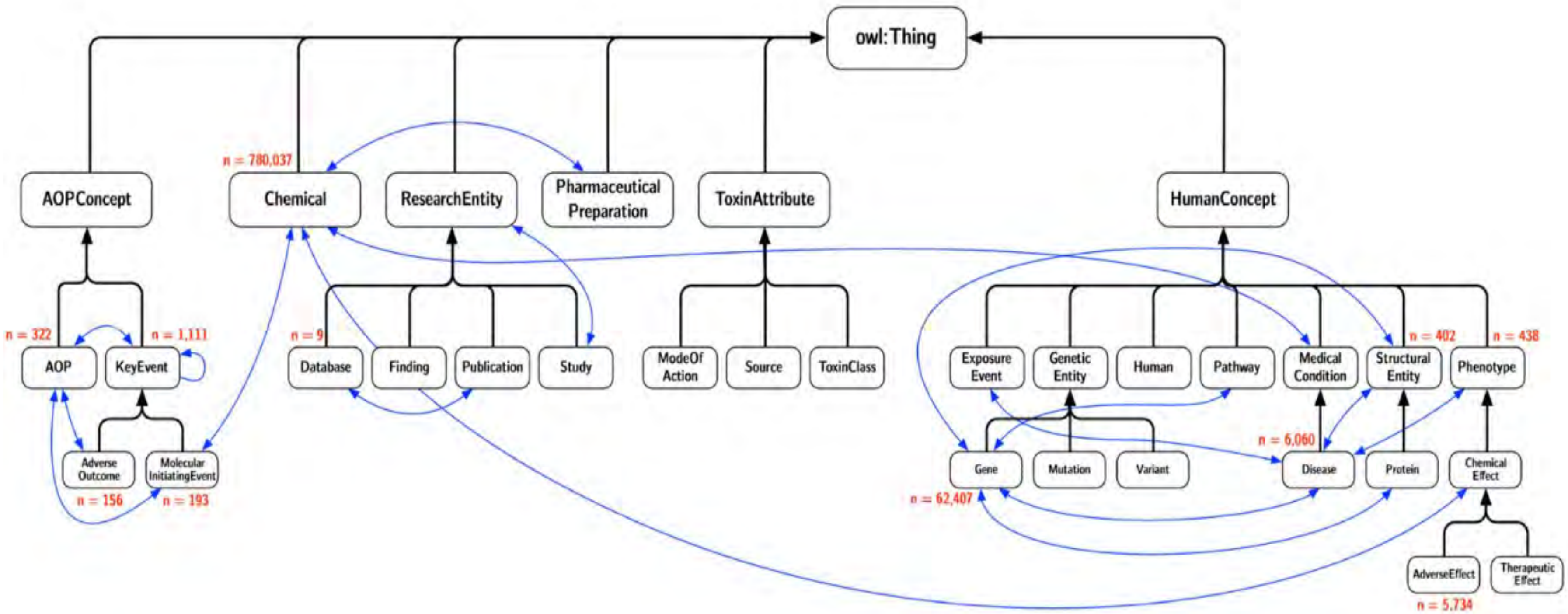
Joseph D. Romano, Yun Hao, Jason H. Moore, and Trevor M. Penning*

Cite This: *Chem. Res. Toxicol.* 2022, 35, 1370–1382

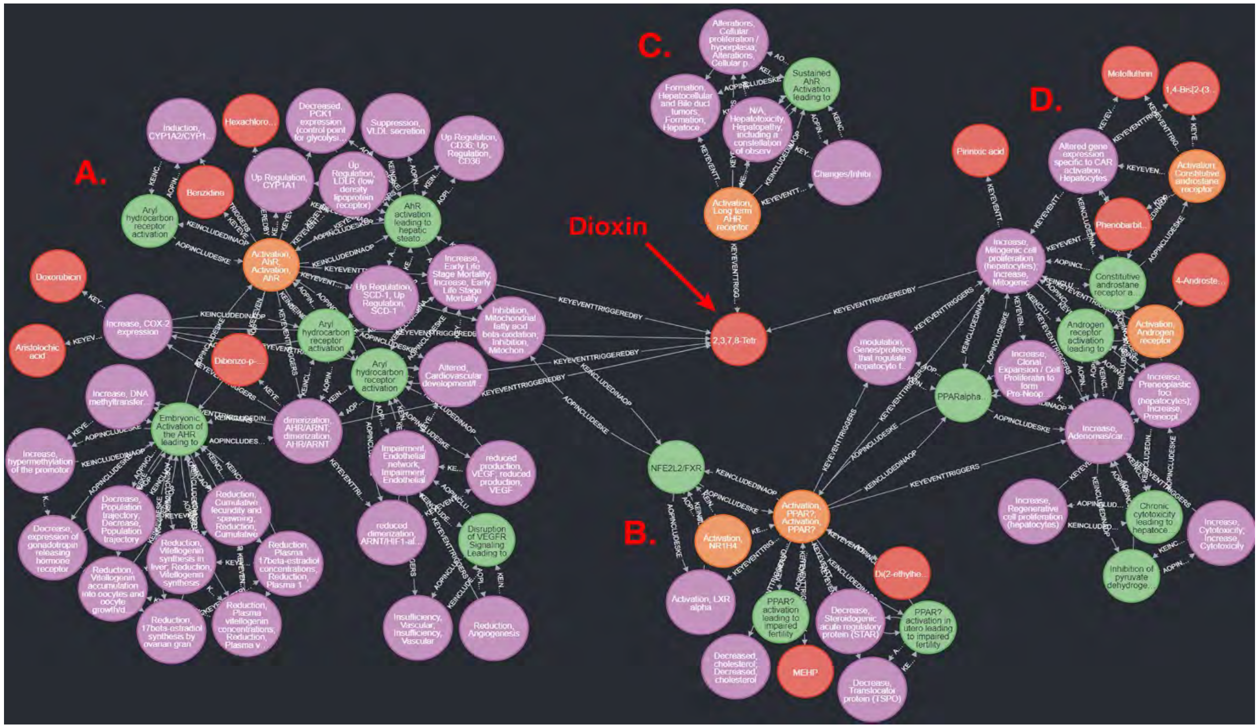
ABSTRACT: ComptoxAI is a new data infrastructure for computational and artificial intelligence research in predictive toxicology. Here, we describe and showcase ComptoxAI's graph-structured knowledge base in the context of three real-world use-cases, demonstrating that it can rapidly answer complex questions about toxicology that are infeasible using previous technologies and data resources. These use-cases each demonstrate a tool for information retrieval from the knowledge base being used to solve a specific task: The "shortest path" module is used to identify mechanistic links between perfluorooctanoic acid (PFOA) exposure and nonalcoholic fatty liver disease; the "expand network"



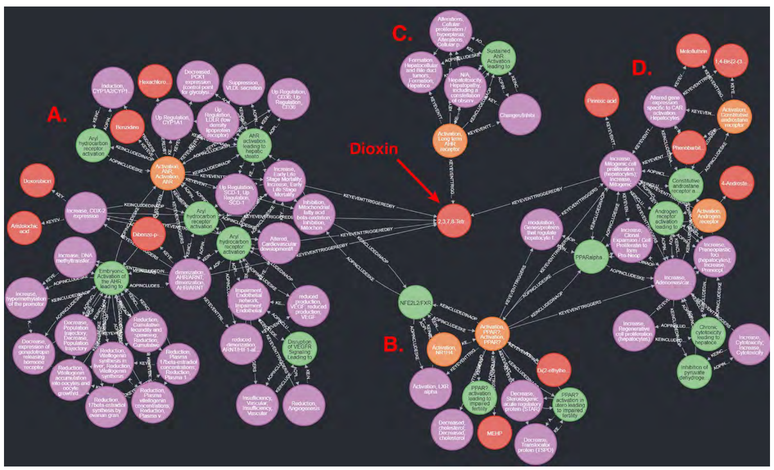
Computational Toxicology Knowledgebase (Comptox.AI)

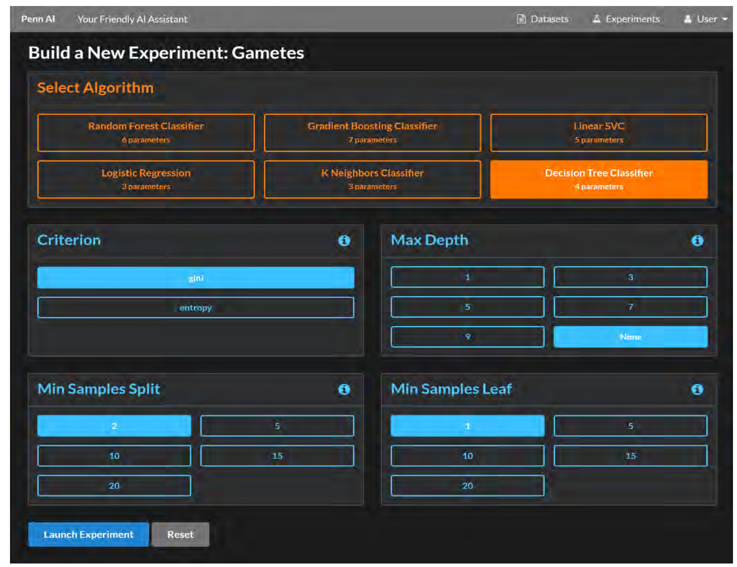


Computational Toxicology Knowledgebase (Comptox.AI)



RESEARCH GOAL: *Knowledge-aware AutoML*





Cedars Sinai


$100
Aliro-ED

4. F

EXTREMETECH

Search Extremetech

SEARCH

Computing Phones Security Gaming Science Space Deep Dives Deals Shop

Job Ads for AI Could Soon Look Like This. Are You Ready?

By Jason H. Moore on November 18, 2020 at 10:07 am

Wanted: Human Assistant to the Artificial Intelligence

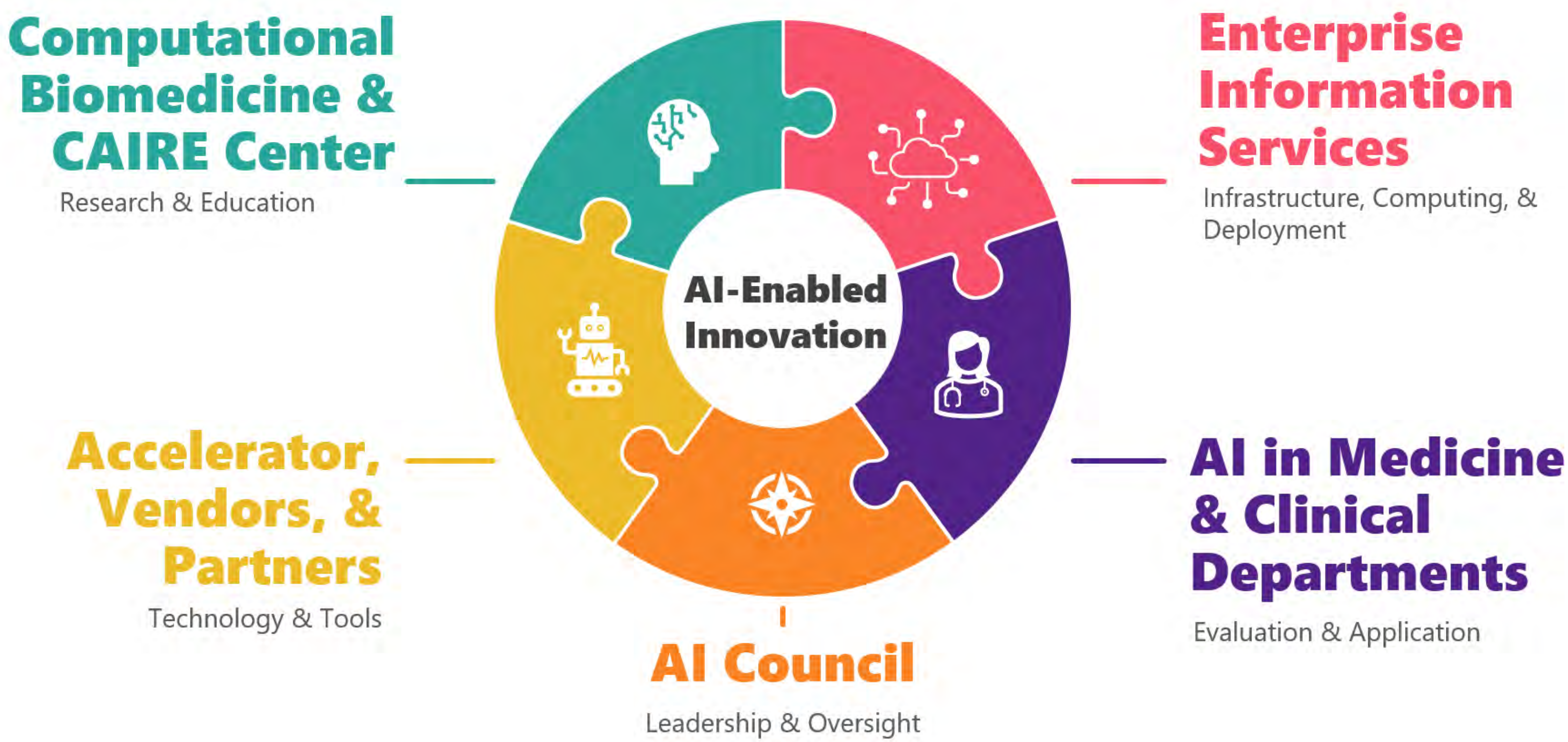
We are seeking junior and mid-level human applicants to serve as data science assistants to our departmental artificial intelligence (AI) in charge of data analytics. Responsibilities include reviewing, interpreting, and providing feedback about analytics results to the AI, and writing summary reports of AI results for human communication. Requires ability to interact with vendors and information technology staff to provide hardware support for the AI. Experience collaborating with computer-based staff a plus. Must have good human-computer interaction skills. Formal training in the ethical treatment of computers and assessment of the fairness and bias of computer-generated results preferred.

euronews.next BIZTECH NEWS MONEY WORK MOBILITY HEALTH HOME OPINION SERIES

WORK

AI prompt engineering: How talking to ChatGPT became the hottest tech job with a six-figure salary

Our Growing AI Ecosystem



Cedars Sinai

Acknowledgements

- AliroAI: Miguel Hernandez, Nick Matsumoto, Jay Moran, Paul Wang
- R01 LM010098, U01 AG066833

- jason.moore@csmc.edu

- https://github.com/epistasislab/aliro
- https://github.com/epistasislab/tpot2
- https://comptox.ai
- https://epistasis.org

Questions?