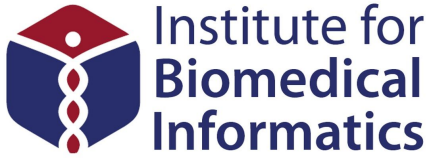


Data science w bioinformatyce

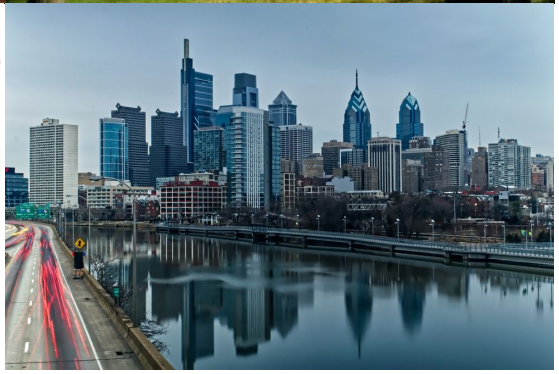
Patryk Orzechowski, Ph.D.
Data Scientist @UPenn
Assistant Professor @AGH

Katedra Automatyki
i Robotyki

Agenda

- Wprowadzenie
- Co to jest Data Science?
- Obszary badawcze
  - Biklasteryzacja i big data
  - Machine Learning i benchmarking
  - Feature selection
  - AutoML
  - Deep Learning
- Wybrane projekty
- Kilka wskazówek ;)
- Podsumowanie

Wprowadzenie

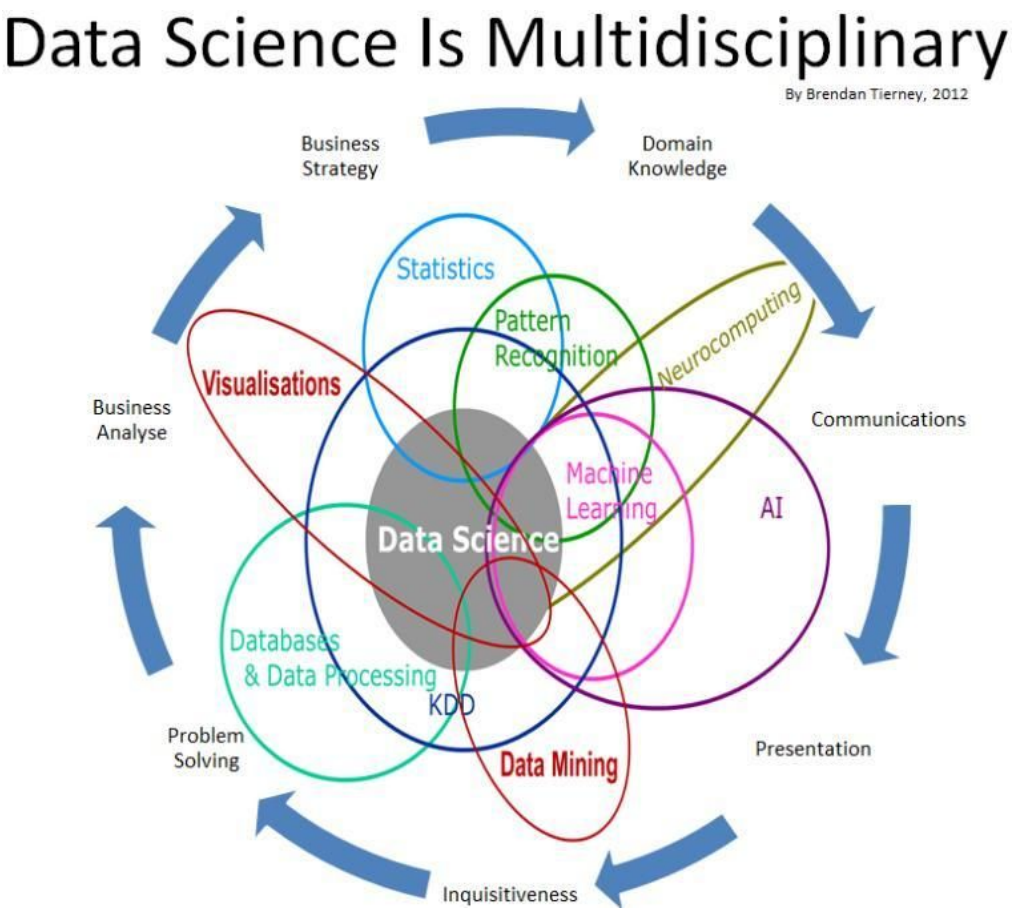


Data Science

> A data scientist is a statistician who lives in San Francisco.
>
> Data Science is statistics on a Mac.
>
> A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

www.quora.com/Data-Science/What-is-the-difference-between-a-data-scientist-and-a-statistician

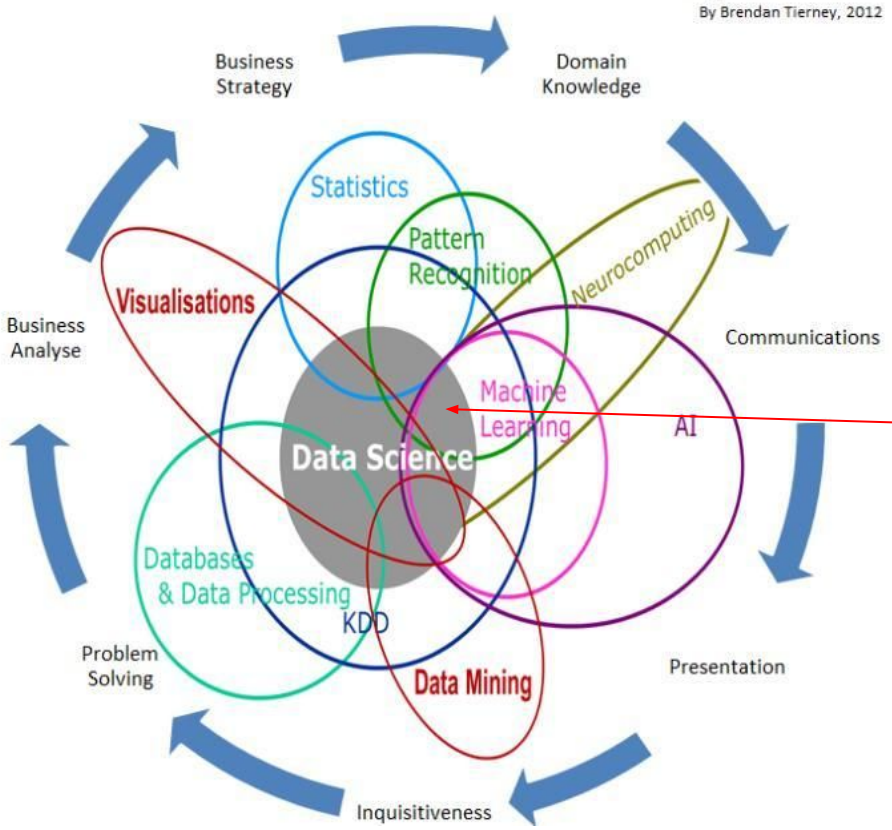
Data Science



Źródło: https://st5.ning.com/topology/rest/1.0/file/get/2808328601

Biklasteryzacja i big data

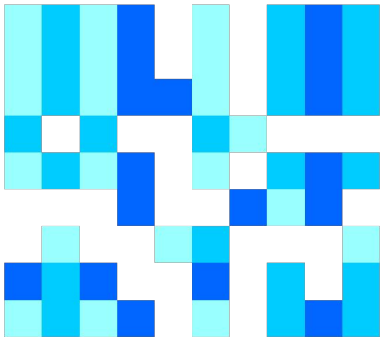
Data Science Is Multidisciplinary



Biclustering

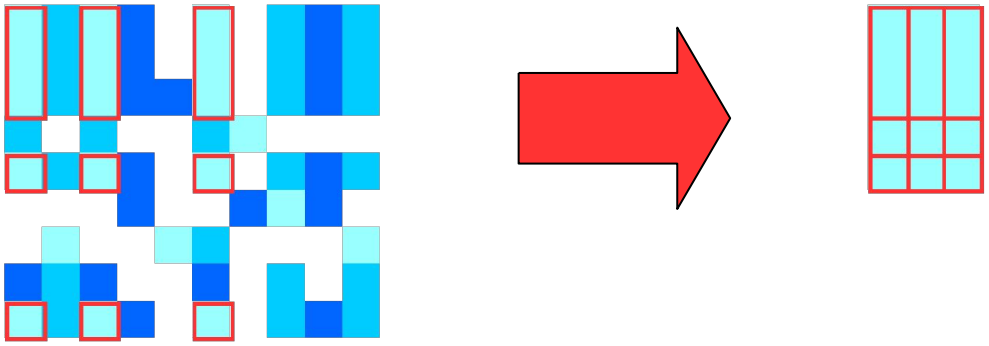
Źródło: https://st5.ning.com/topology/rest/1.0/file/get/2808328601

Biklasteryzacja



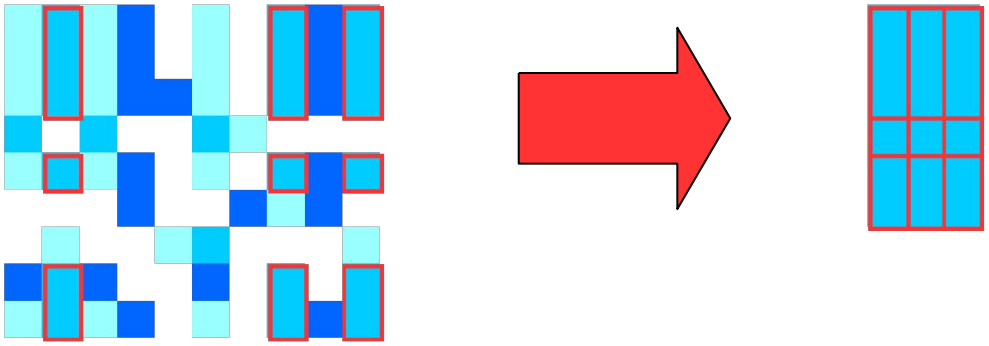
Biklasteryzacja (ang. biclustering, co-clustering, two-mode clustering) to nienadzorowana technika uczenia maszynowego polegająca na równoczesnym grupowaniu wierszy i kolumn.

Biklasteryzacja



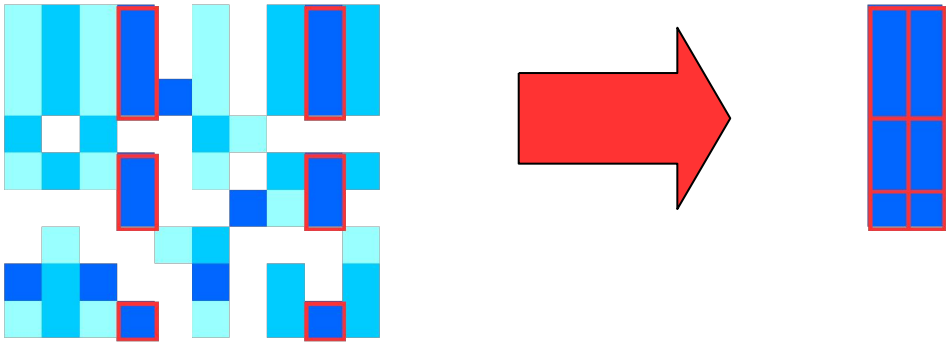
Biklasteryzacja (ang. biclustering, co-clustering, two-mode clustering) to nienadzorowana technika uczenia maszynowego polegająca na równoczesnym grupowaniu wierszy i kolumn.

Biklasteryzacja



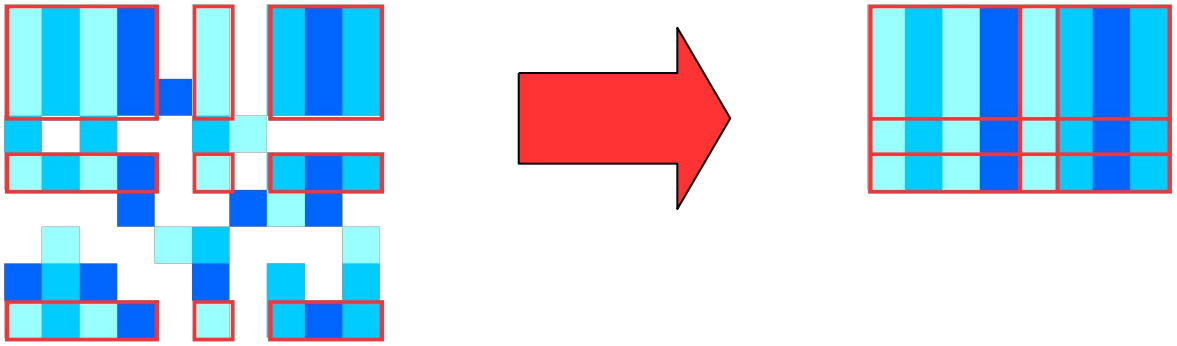
Biklasteryzacja (ang. biclustering, co-clustering, two-mode clustering) to nienadzorowana technika uczenia maszynowego polegająca na równoczesnym grupowaniu wierszy i kolumn.

Biklasteryzacja



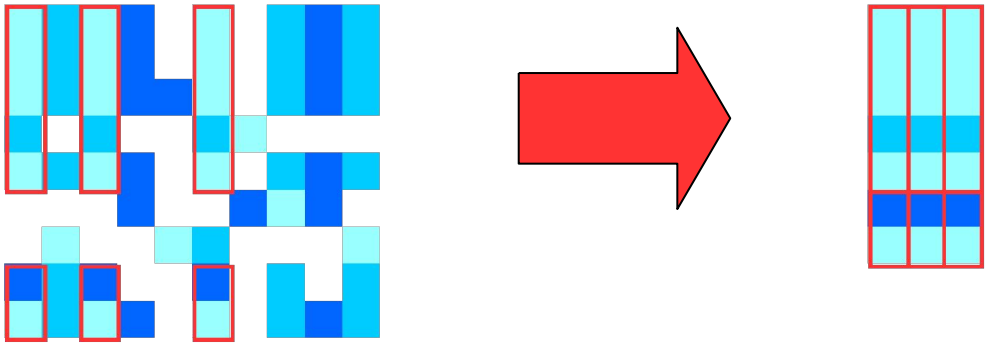
Biklasteryzacja (ang. biclustering, co-clustering, two-mode clustering) to nienadzorowana technika uczenia maszynowego polegająca na równoczesnym grupowaniu wierszy i kolumn.

Biklasteryzacja



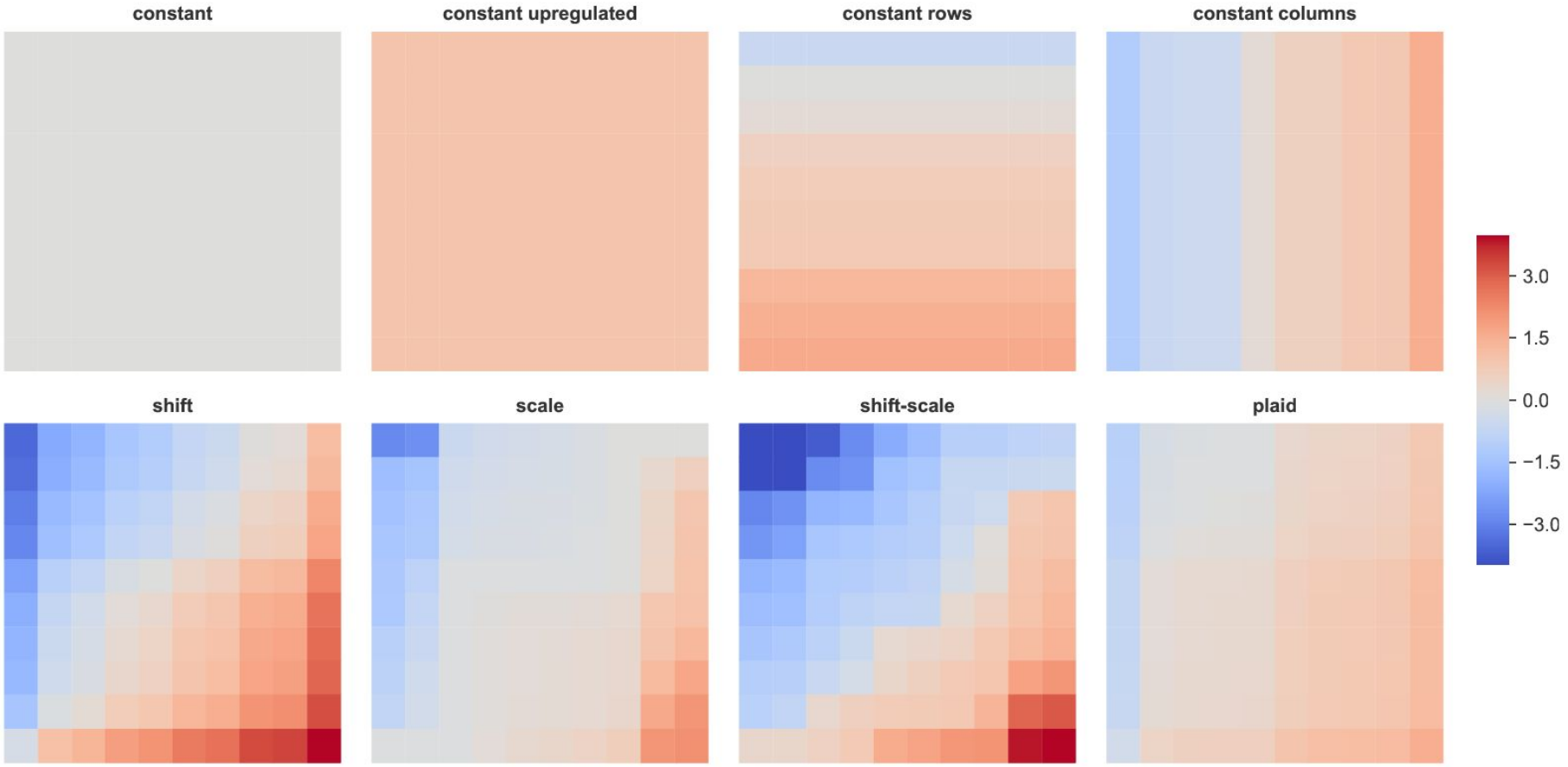
Rezultatem biklasteryzacji jest zbiór biklastrów. Biklaster to podzbiór wierszy i podzbiór kolumn, w których wiersze “zachowują się” podobnie do siebie względem kolumn (lub vice-versa).

Biklasteryzacja

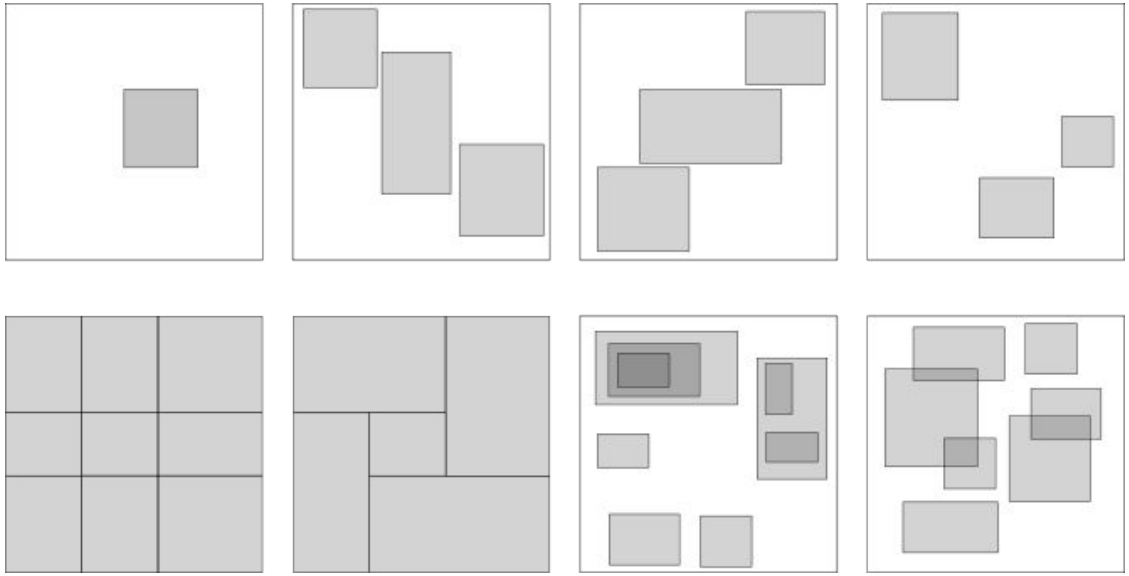


Rezultatem biklasteryzacji jest zbiór biklastrów. Biklaster to podzbiór wierszy i podzbiór kolumn, w których wiersze “zachowują się” podobnie do siebie względem kolumn (lub vice-versa).

Typy biklastrów

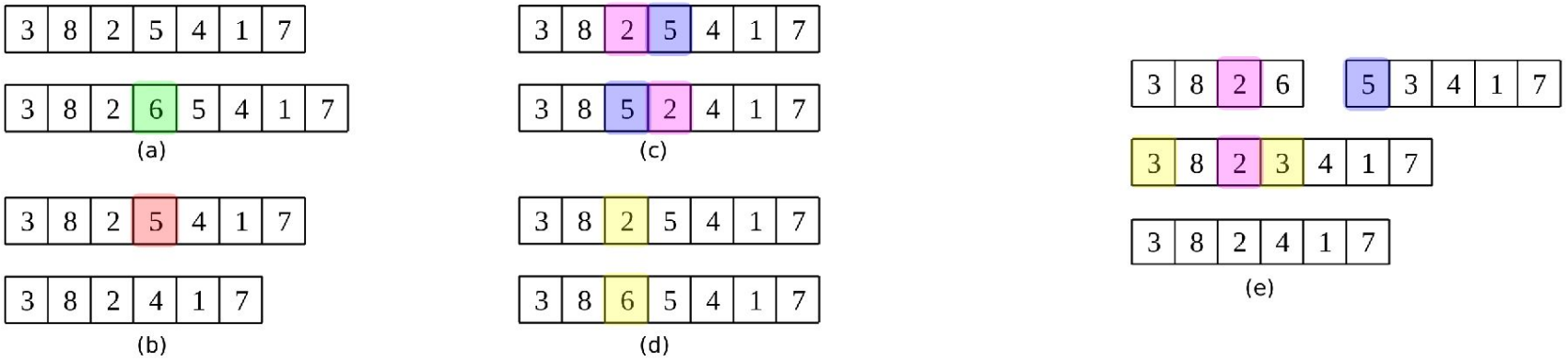


Struktura biklastrów

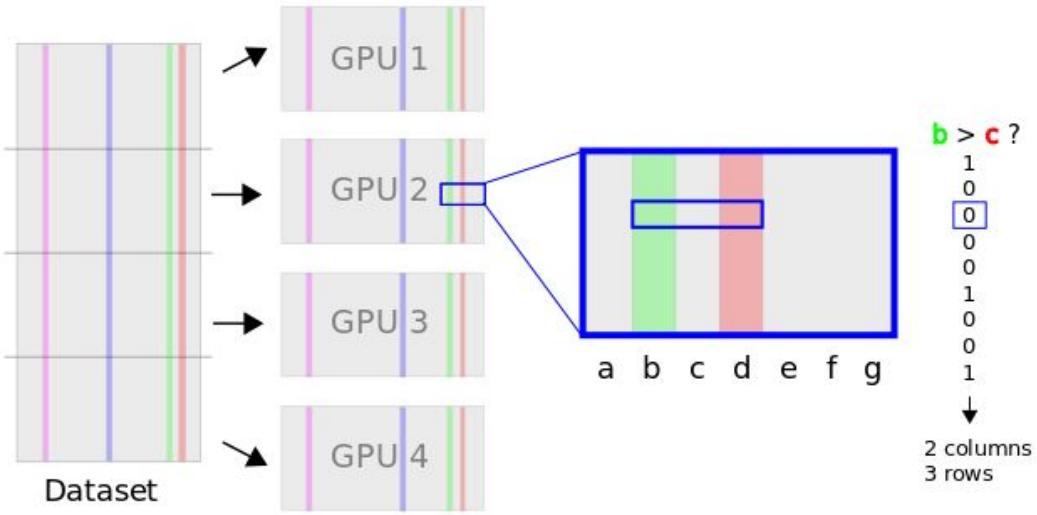


EBIC - biclustering big data

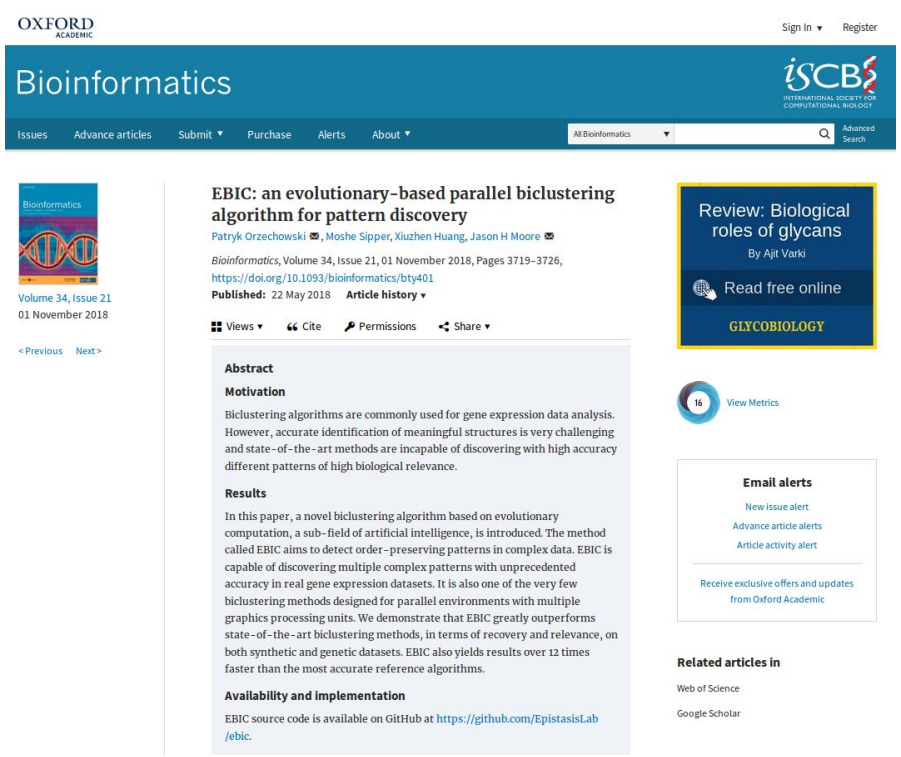
- **E**volutionary search-based **Bic**lustering (EBIC)
- działa na klastrze kart graficznych
- C++11/CUDA 8
- różne strategie ewolucyjne
- proste operacje genetyczne:



EBIC - biclustering big data



EBIC - publikacje



OXFORD ACADEMIC

Sign In ▾ Register

Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues Advance articles Submit ▾ Purchase Alerts About ▾ All Bioinformatics Advanced Search

Volume 34, Issue 21
01 November 2018

< Previous Next >

EBIC: an evolutionary-based parallel biclustering algorithm for pattern discovery

Patryk Orzechowski ✉, Moshe Sipper, Xiuzhen Huang, Jason H Moore ✉

Bioinformatics, Volume 34, Issue 21, 01 November 2018, Pages 3719–3726, https://doi.org/10.1093/bioinformatics/bty401

Published: 22 May 2018 **Article history** ▾

Views ▾ Cite Permissions Share ▾

Abstract

Motivation

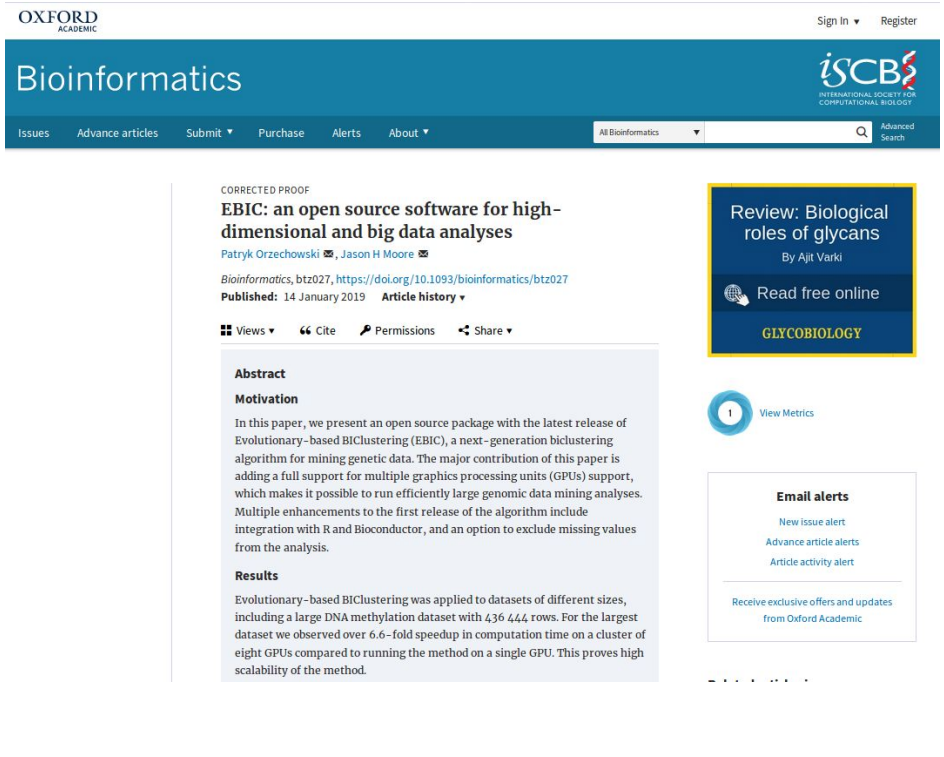
Biclustering algorithms are commonly used for gene expression data analysis. However, accurate identification of meaningful structures is very challenging and state-of-the-art methods are incapable of discovering with high accuracy different patterns of high biological relevance.

Results

In this paper, a novel biclustering algorithm based on evolutionary computation, a sub-field of artificial intelligence, is introduced. The method called EBIC aims to detect order-preserving patterns in complex data. EBIC is capable of discovering multiple complex patterns with unprecedented accuracy in real gene expression datasets. It is also one of the very few biclustering methods designed for parallel environments with multiple graphics processing units. We demonstrate that EBIC greatly outperforms state-of-the-art biclustering methods, in terms of recovery and relevance, on both synthetic and genetic datasets. EBIC also yields results over 12 times faster than the most accurate reference algorithms.

Availability and implementation

EBIC source code is available on GitHub at https://github.com/EpistasisLab/ebic.



OXFORD ACADEMIC

Sign In ▾ Register

Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues Advance articles Submit ▾ Purchase Alerts About ▾ All Bioinformatics Advanced Search

CORRECTED PROOF

EBIC: an open source software for high-dimensional and big data analyses

Patryk Orzechowski ✉, Jason H Moore ✉

Bioinformatics, btz027, https://doi.org/10.1093/bioinformatics/btz027

Published: 14 January 2019 **Article history** ▾

Views ▾ Cite Permissions Share ▾

Abstract

Motivation

In this paper, we present an open source package with the latest release of Evolutionary-based BIClustering (EBIC), a next-generation biclustering algorithm for mining genetic data. The major contribution of this paper is adding a full support for multiple graphics processing units (GPUs) support, which makes it possible to run efficiently large genomic data mining analyses. Multiple enhancements to the first release of the algorithm include integration with R and Bioconductor, and an option to exclude missing values from the analysis.

Results

Evolutionary-based BIClustering was applied to datasets of different sizes, including a large DNA methylation dataset with 436 444 rows. For the largest dataset we observed over 6.6-fold speedup in computation time on a cluster of eight GPUs compared to running the method on a single GPU. This proves high scalability of the method.

Biklasteryzacja – wyzwania

OXFORD

(GIGA)n SCIENCE

GigaScience, 2017, 1–4

doi: xx.xxxx/xxxx
Manuscript in Preparation
Commentary

COMMENTARY

Scalable biclustering – the future of big data exploration?

Patryk Orzechowski[1,2*], Krzysztof Boryczko[3] and Jason H. Moore[1*]

[1]Institute for Biomedical Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, USA and [2]Department of Automatics and Robotics, AGH University of Science and Technology, al. A. Mickiewicza 30, Kraków, 30-059, Poland and [3]Department of Computer Science, AGH University of Science and Technology, al. A. Mickiewicza 30, Kraków, 30-059, Poland

*Corresponding author: patryk.orzechowski@gmail.com

Abstract

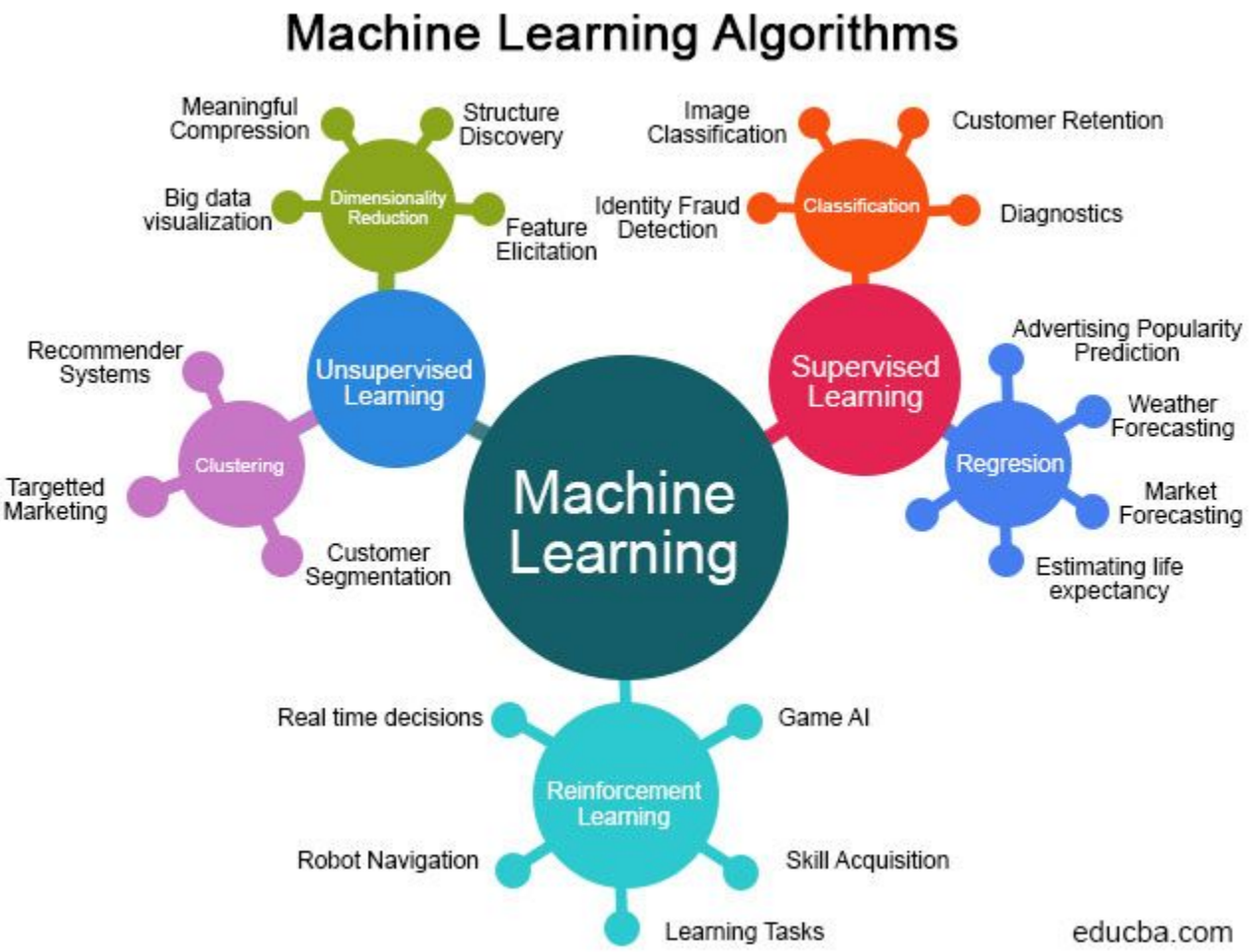
Biclustering is a technique of discovering local similarities within data. For many years the complexity of the methods and parallelization issues limited its application to big data problems. With development of novel scalable methods, biclustering has finally started to close this gap. In this paper we discuss caveats of biclustering, present its current challenges and guidelines for practitioners. We also try to explain why biclustering may soon become one of the standards for big data analytics.

Key words: biclustering, co-clustering, data mining, big data, parallel algorithms, disease subtype identification, biomarker detection, gene-drug interaction, precision medicine

Machine Learning

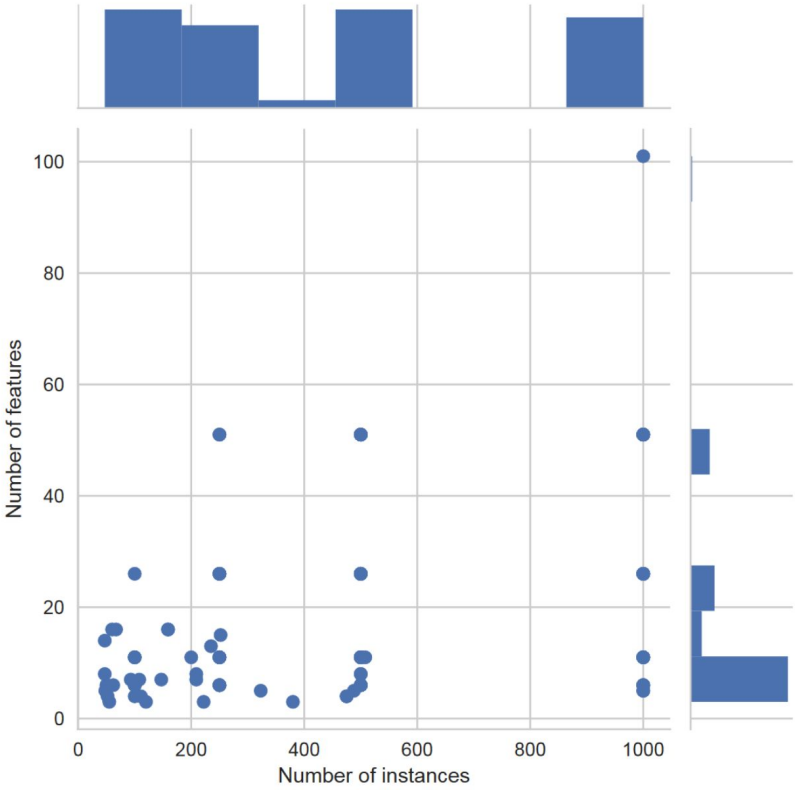
Benchmarking

Machine Learning



Źródło:https://www.educba.com/machine-learning-algorithms/

Penn ML Benchmarks (PMLB)



EpistasisLab / penn-ml-benchmarks

Used by 19 | Watch 26 | Star 488 | Fork 82

Code | Issues 11 | Pull requests 0 | Actions | Projects 1 | Wiki | Security | Insights

PMLB: A large, curated repository of benchmark datasets for evaluating supervised machine learning algorithms.
https://biodatamining.biomedcentral.c...

33 commits | 2 branches | 0 packages | 0 releases | Fetching contributors | MIT

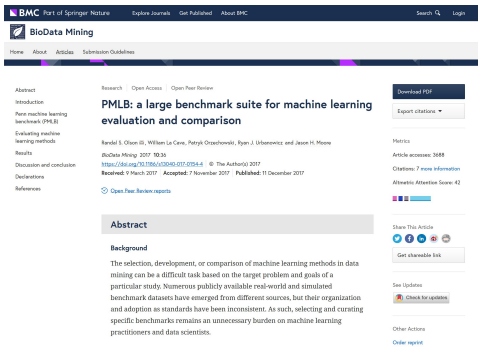
Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

rhiever Merge pull request #18 from weixuanfu/patch-1 ... Latest commit ec90219 Feb 13, 2018

datasets	Update README.md	Feb 13, 2018
pmlb	Code clean up + docs update for release	Dec 28, 2017
.gitignore	Initial commit	Nov 11, 2016
LICENSE	Initial commit	Nov 11, 2016
MANIFEST.in	Add manifest so the Python package doesn't include all datasets	Dec 5, 2016
README.md	Add example PMLB usage for benchmarking to README	Dec 29, 2017
setup.py	Add wrapper for pmlb	Dec 5, 2016

README.md

Penn Machine Learning Benchmarks



BMC Part of Springer Nature

BioData Mining

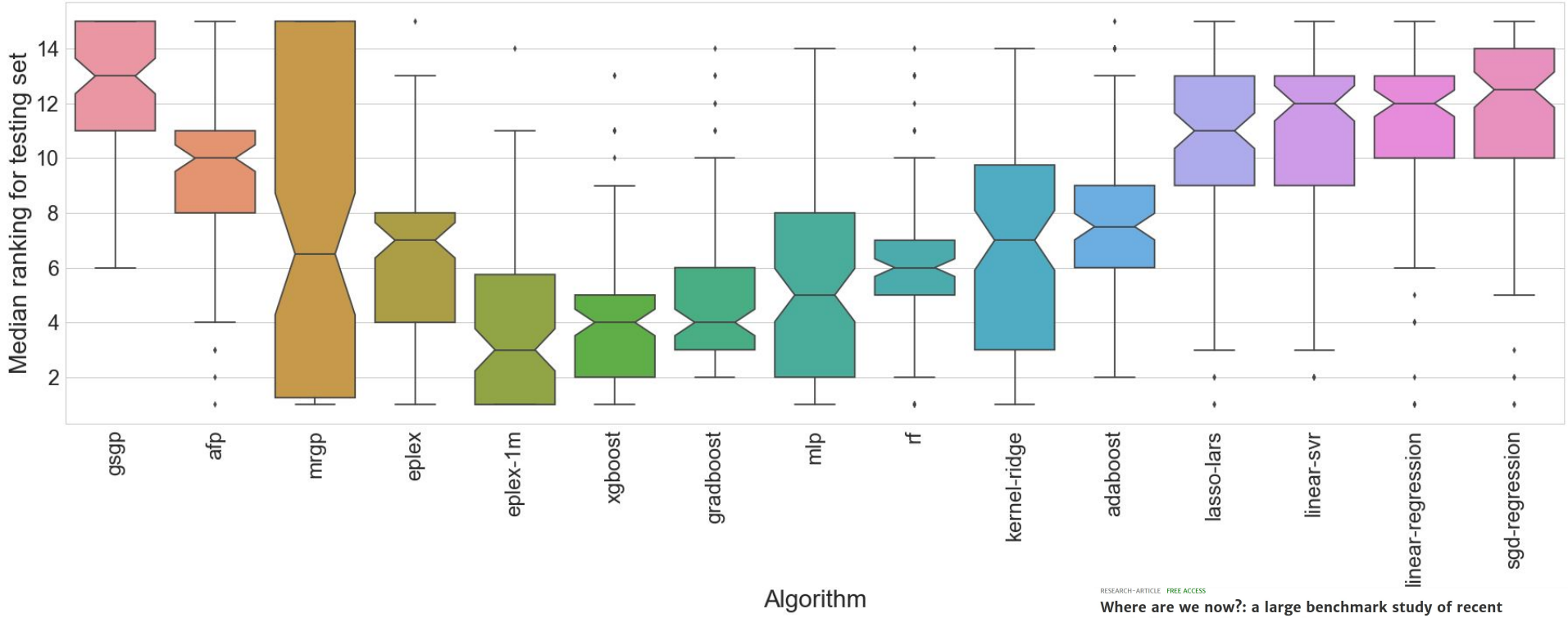
PMLB: a large benchmark suite for machine learning evaluation and comparison

Abstract

Background

The selection, development, or comparison of machine learning methods in data mining can be a difficult task based on the target problem and goals of a particular study. Numerous publicly available real-world and simulated benchmark datasets have emerged from different sources, but their organization and adoption as standards have been inconsistent. As such, selecting and curating specific benchmarks remains an unnecessary burden on machine learning practitioners and data scientists.

Porównanie metod regresji



RESEARCH-ARTICLE FREE ACCESS

Where are we now?: a large benchmark study of recent symbolic regression methods

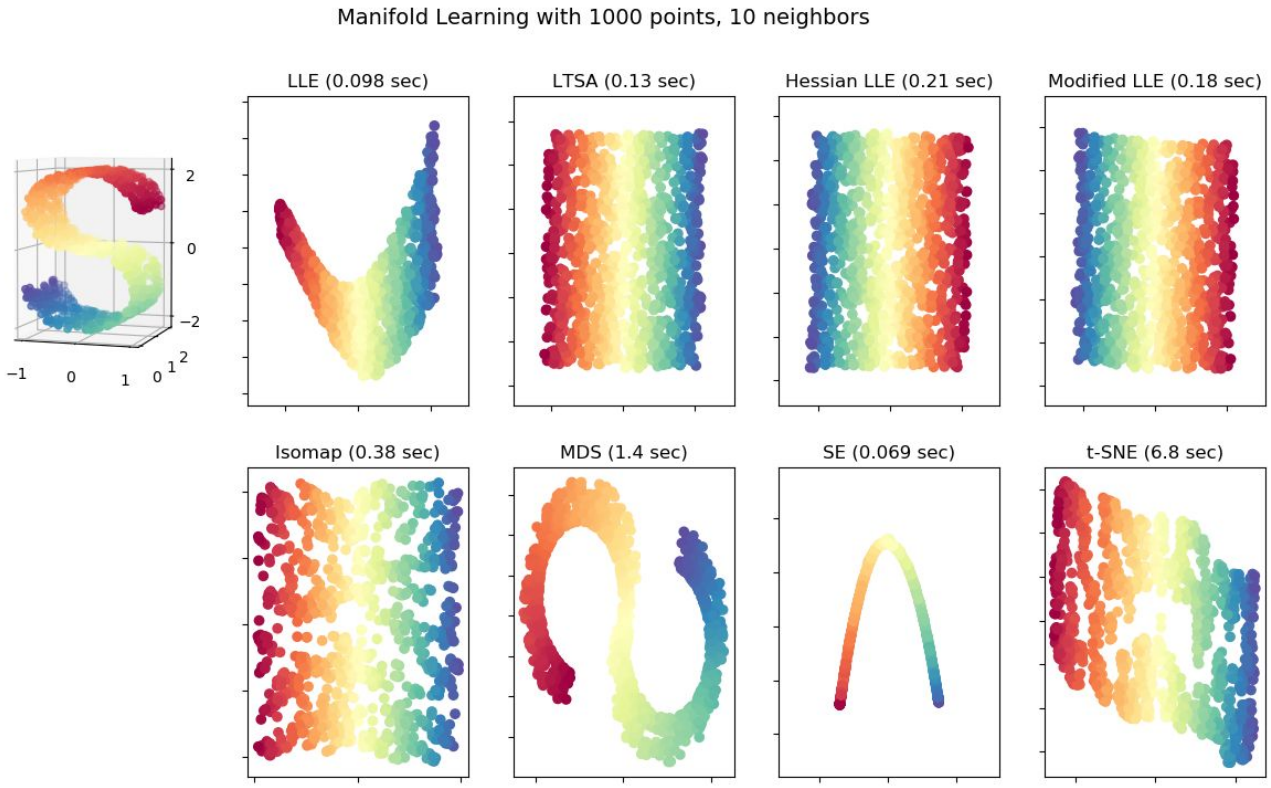
Authors: Patryk Orzechowski, William La Cava, Jason H. Moore Authors Info & Affiliations

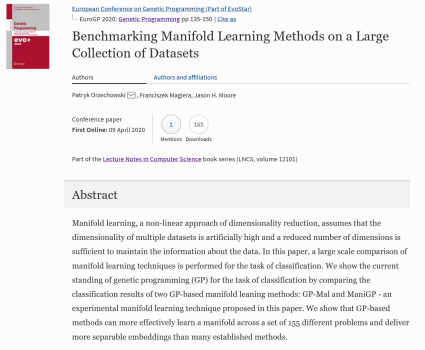
Publication: GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference • July 2018 • Pages 1183–1190 • https://doi.org/10.1145/3205455.3205539

ABSTRACT

In this paper we provide a broad benchmarking of recent genetic programming approaches to symbolic regression in the context of state of the art machine learning approaches. We use a set of nearly 100 regression benchmark problems culled from open source repositories across the web. We conduct a rigorous benchmarking of four recent symbolic regression approaches as well as nine machine learning approaches from scikit-learn. The results suggest that symbolic regression performs strongly compared to state-of-the-art gradient boosting

Manifold Learning





European Conference on Genetic Programming (Part of EvoStar)

EuroGP 2020: Genetic Programming pp 135-150 | Cite as

Benchmarking Manifold Learning Methods on a Large Collection of Datasets

Authors | Authors and affiliations

Patryk Orzechowski, Franciszek Magiera, Jason H. Moore

Conference paper

First Online: 09 April 2020

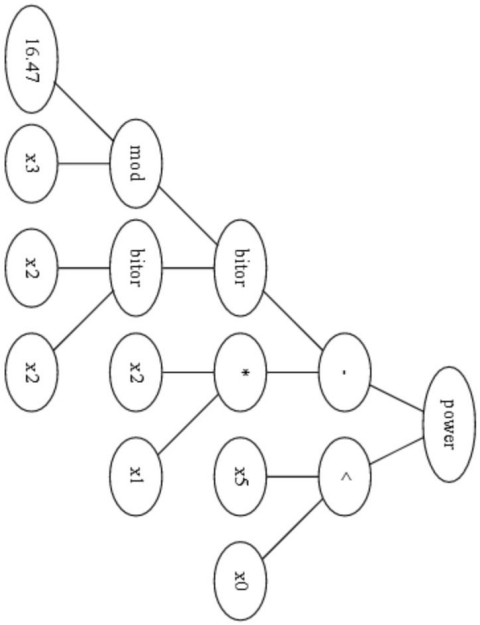
1 Mentions | 165 Downloads

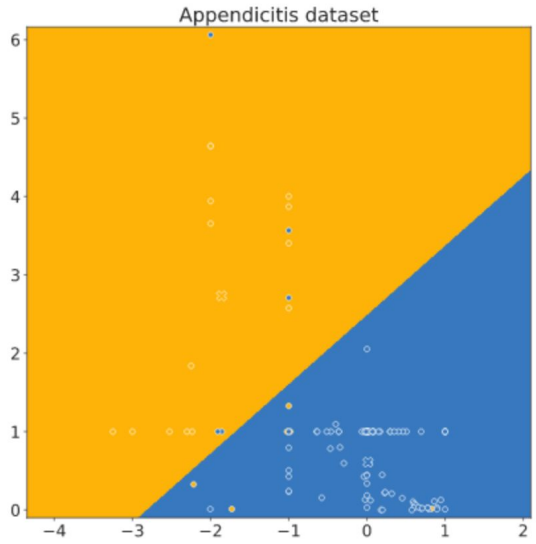
Part of the Lecture Notes in Computer Science book series (LNCS, volume 12101)

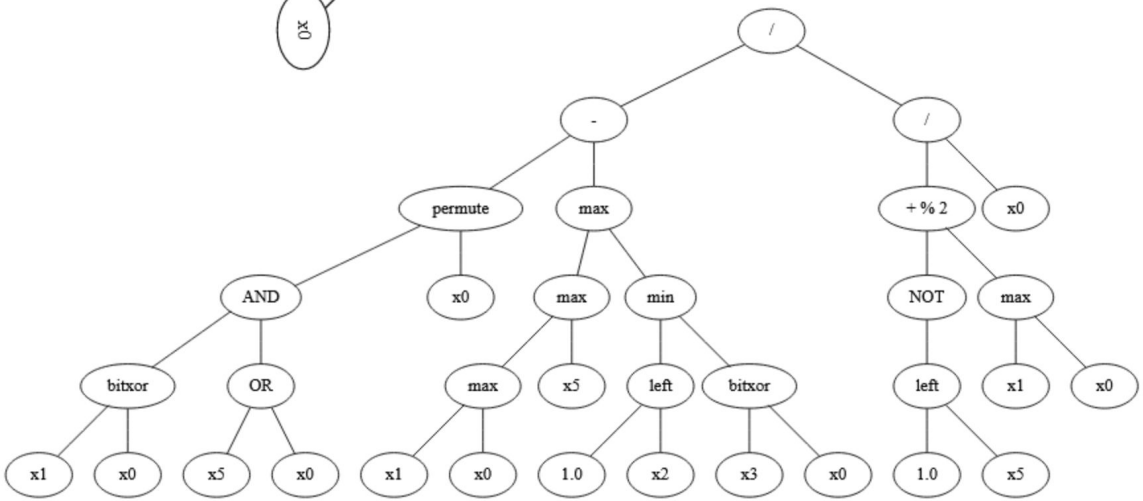
Abstract

Manifold learning, a non-linear approach of dimensionality reduction, assumes that the dimensionality of multiple datasets is artificially high and a reduced number of dimensions is sufficient to maintain the information about the data. In this paper, a large scale comparison of manifold learning techniques is performed for the task of classification. We show the current standing of genetic programming (GP) for the task of classification by comparing the classification results of two GP-based manifold leaning methods: GP-Mal and ManiGP - an experimental manifold learning technique proposed in this paper. We show that GP-based methods can more effectively learn a manifold across a set of 155 different problems and deliver more separable embeddings than many established methods.

Genetic Programming

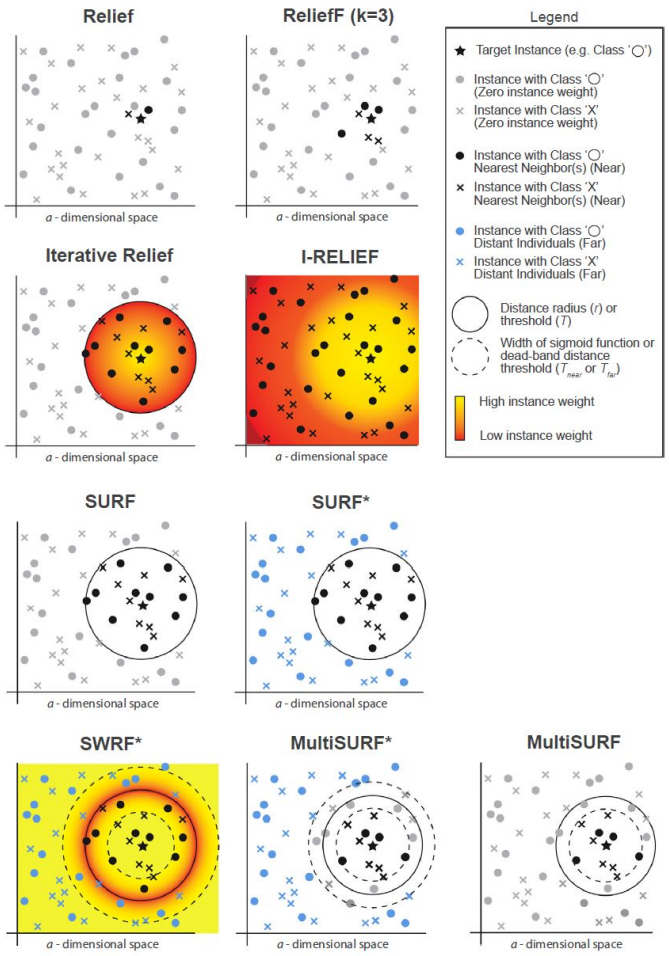






Feature selection

Scikit-rebate



Why GitHub? Enterprise Explore Marketplace Pricing Search Sign in Sign up

EpistasisLab / scikit-rebate

Watch 17 Star 178 Fork 28

Code Issues 11 Pull requests 1 Projects 0 Insights

Join GitHub today

GitHub is home to over 31 million developers working together to host and review code, manage projects, and build software together.

Sign up

Dismiss

A scikit-learn-compatible Python implementation of ReBATE, a suite of Relief-based feature selection algorithms for Machine Learning.
https://EpistasisLab.github.io/scikit...

feature-selection data-science python

222 commits 3 branches 0 releases Fetching contributors MIT

Branch: master New pull request Find File Clone or download

Fetching latest commit...

.settings		
ci		
data		
docs	Minor update to installing documentation	May 22, 2018

ELSEVIER

Journal of Biomedical Informatics
Volume 85, September 2018, Pages 189-203

Relief-based feature selection: Introduction and review

Ryan J. Urbanowicz [a], Melissa Meeker [b], William La Cava [a], Randal S. Olson [a], Jason H. Moore [a]

Show more

https://doi.org/10.1016/j.jbi.2018.07.014 Get rights and content

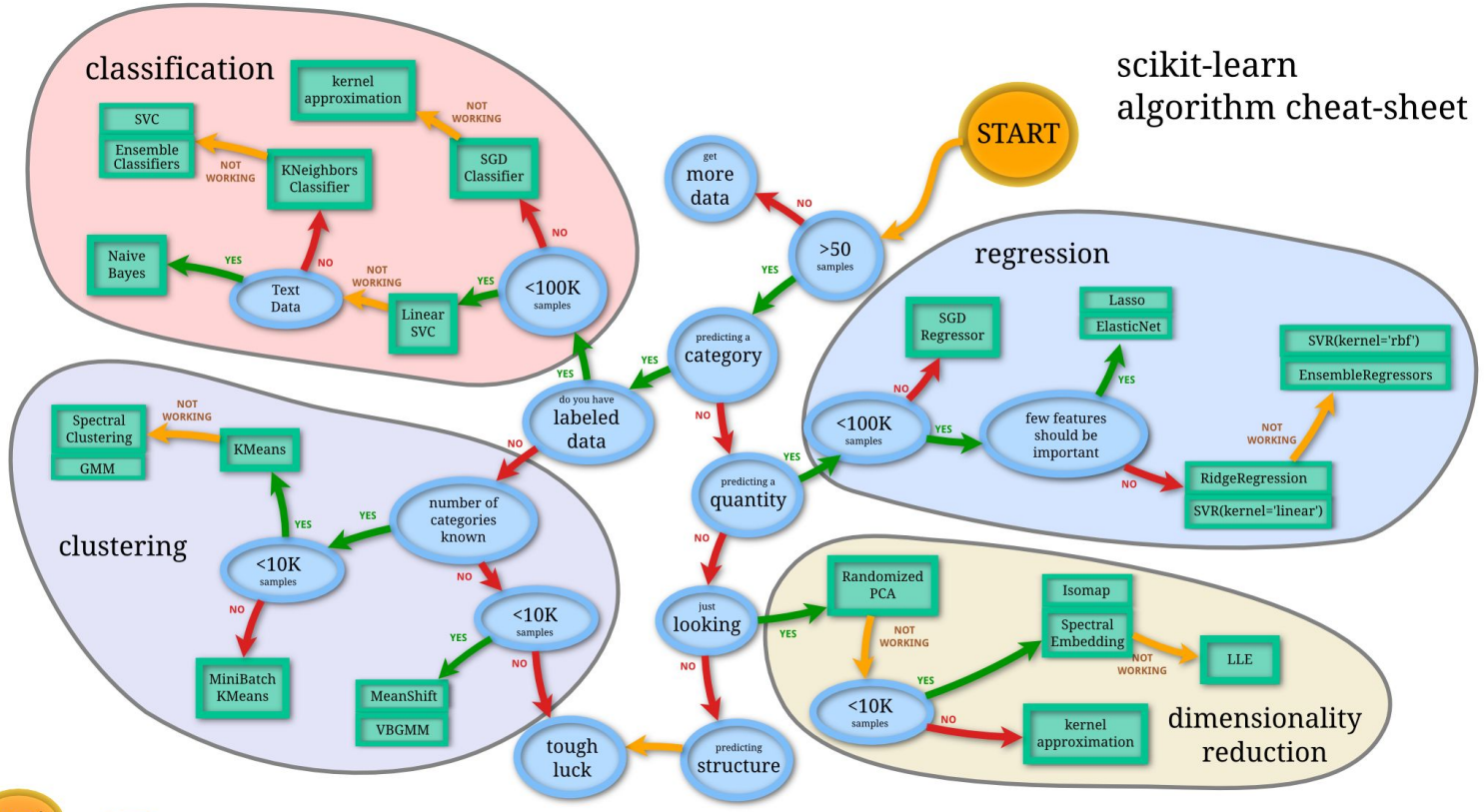
Under an Elsevier user license open archive

Highlights

- Relief-based feature selection methods (RBAs) are reviewed in detailed context.
- RBAs can detect interactions without examining pairwise combinations.
- Iterative RBAs have been developed to scale them up to very large feature spaces.

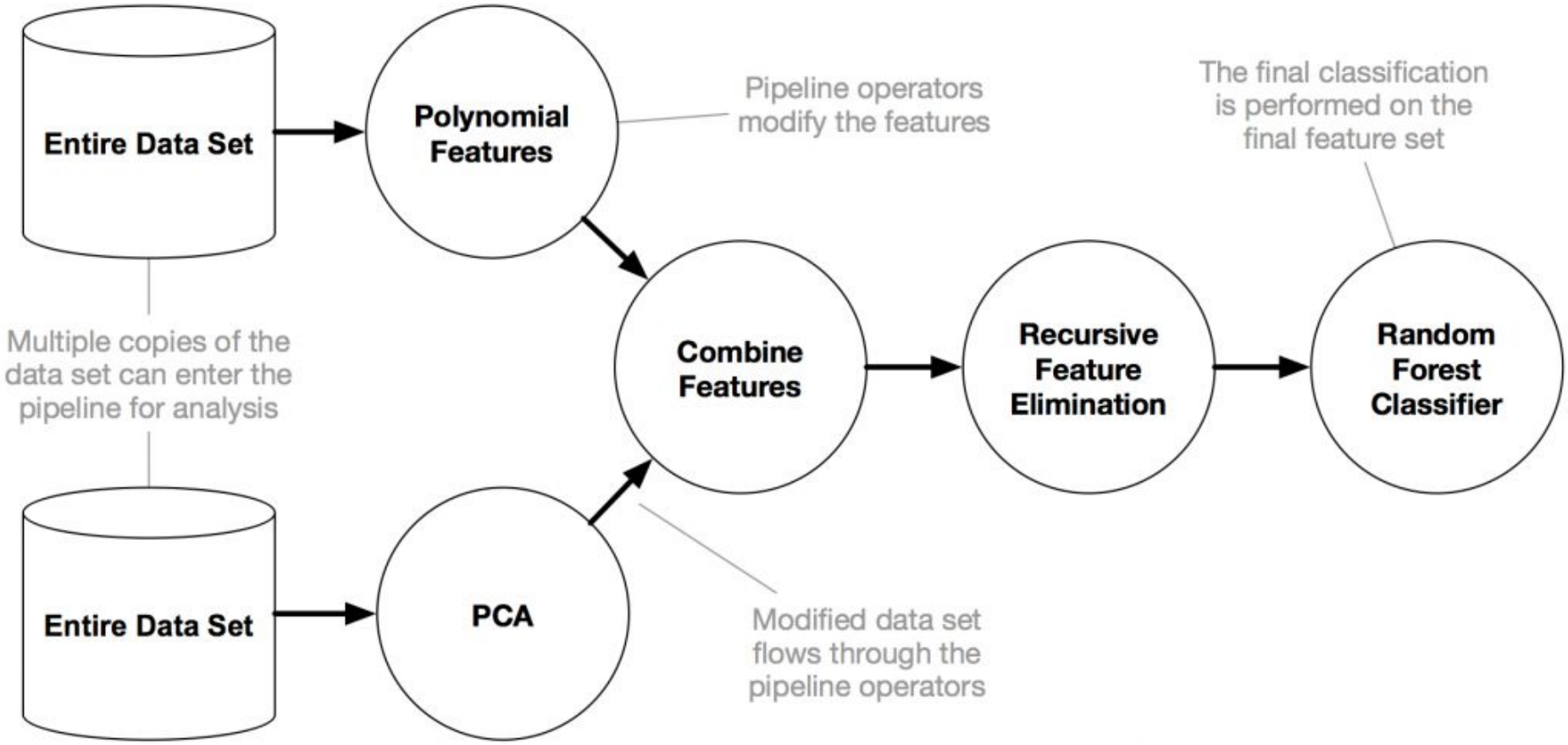
AutoML

AutoML



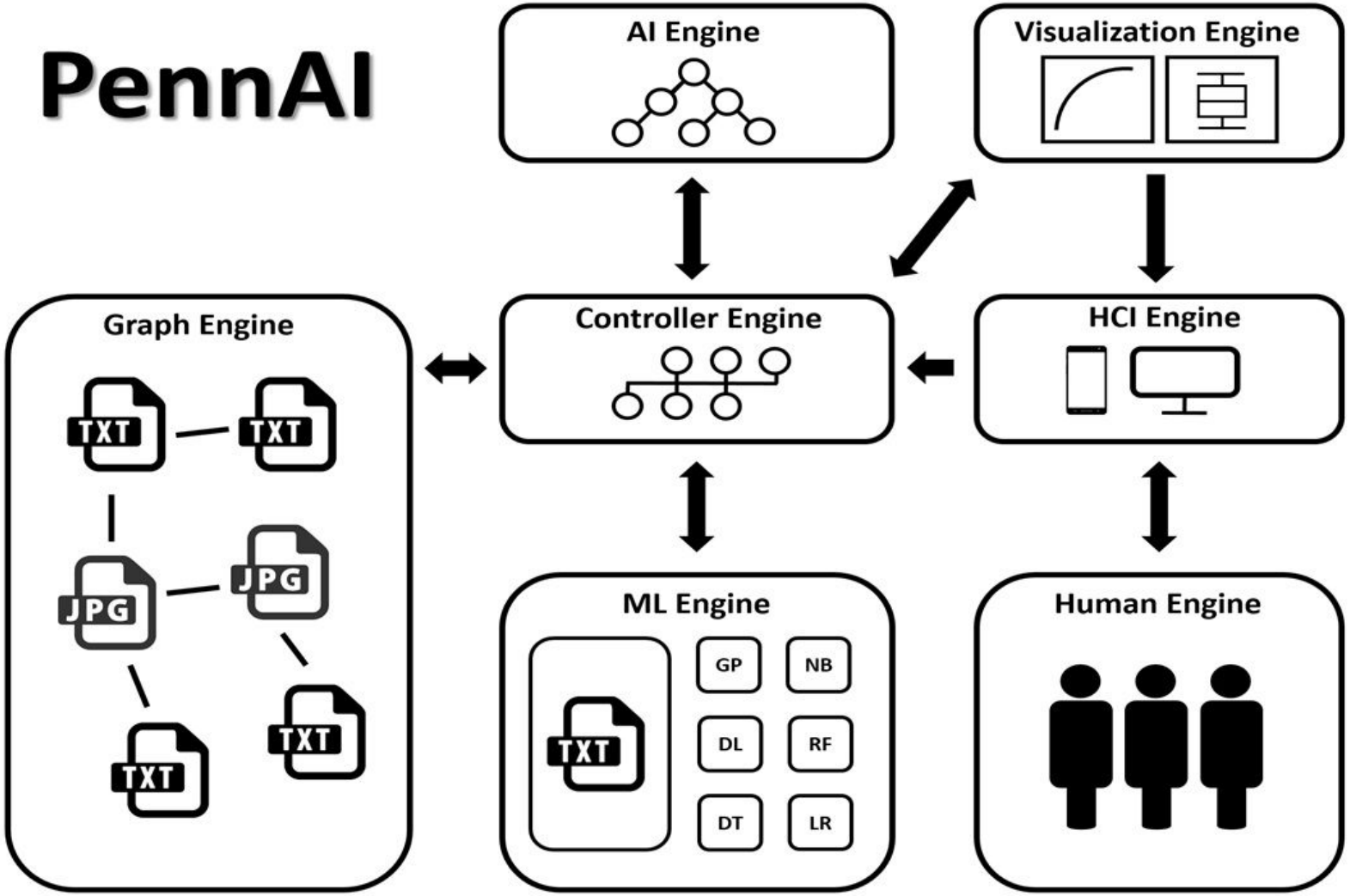
Źródło: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

AutoML - TPOT



Źródło: https://github.com/EpistasisLab/tpot

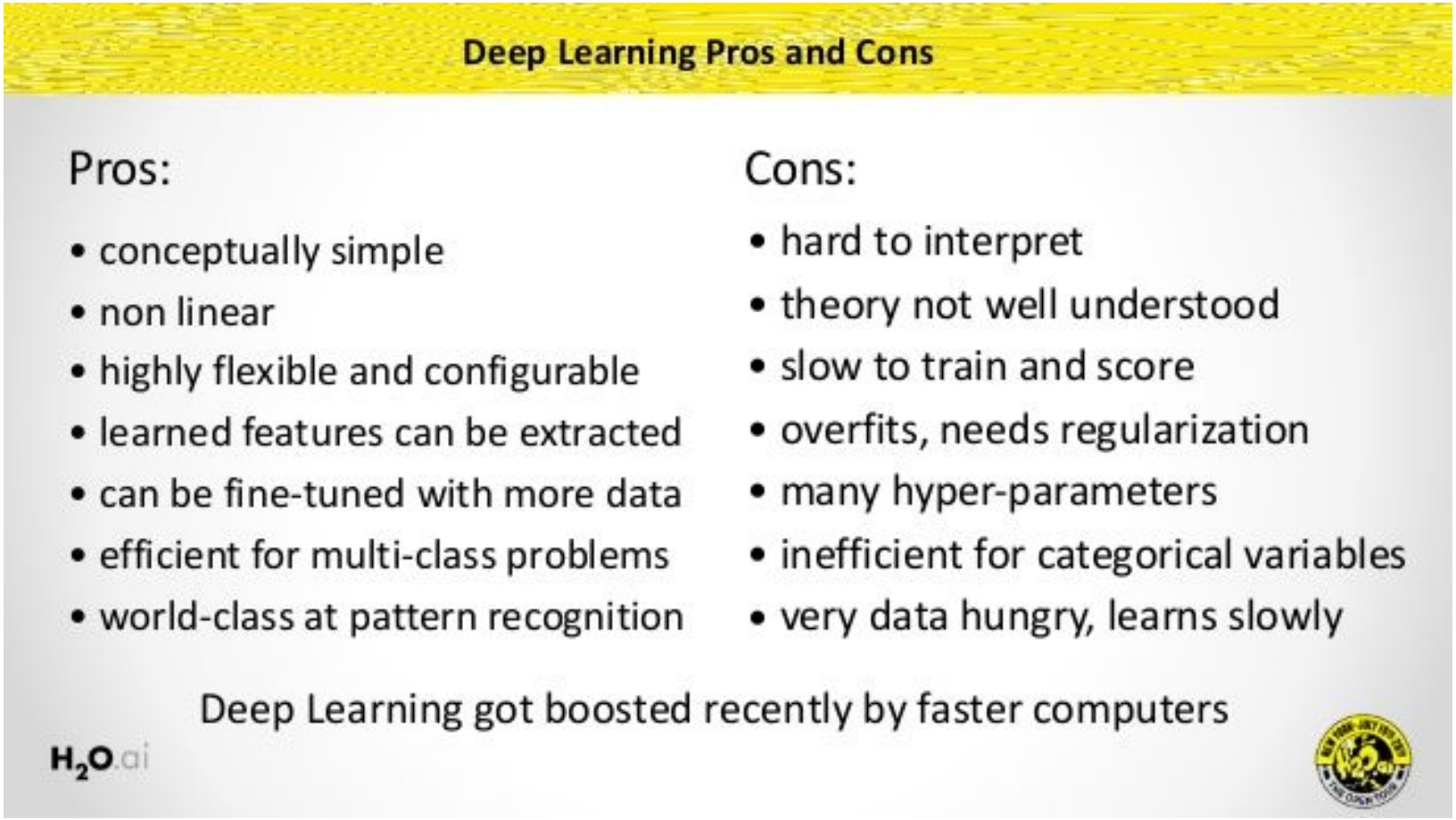
AutoML - PennAI



Źródło: http://pennai.org/

Deep Learning

Deep Learning



Źródło: https://www.slideshare.net/0xdata/transform-your-business-with-ai-deep-learning-and-machine-learning

Deep learning - przegląd

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS

Search

Advanced Search

New Results

Previous Next

Opportunities and obstacles for deep learning in biology and medicine

Posted January 19, 2018.

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H.S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, Casey S. Greene

doi: https://doi.org/10.1101/142760

Abstract | Full Text | Info/History | Metrics | Preview PDF

Download PDF

Supplementary Material

Email

Share

Citation Tools

https://www.fac

Subject Area

Bioinformatics

Abstract

Deep learning, which describes a class of machine learning algorithms, has recently showed impressive results across a variety of domains. Biology and medicine are data rich, but the data are complex and often ill-understood. Problems of this nature may be particularly well-suited to deep learning techniques. We examine applications of deep learning to a variety of biomedical problems—patient classification, fundamental biological processes, and treatment of patients—and discuss whether deep learning will transform these tasks or if the biomedical sphere poses unique challenges. We find that deep learning has yet to revolutionize or definitively resolve any of these problems, but promising advances have been made on the prior state of the art. Even when improvement over a previous baseline has been modest, we have seen signs that deep learning methods may speed or aid human investigation. More work is needed to address concerns related to interpretability and how to best model each problem. Furthermore, the limited amount of labeled data for training presents problems in some domains, as do legal and privacy constraints on work with sensitive health records. Nonetheless, we foresee deep learning powering changes at both bench and bedside with the potential to transform several areas of biology and medicine.

Subject Areas

All Articles

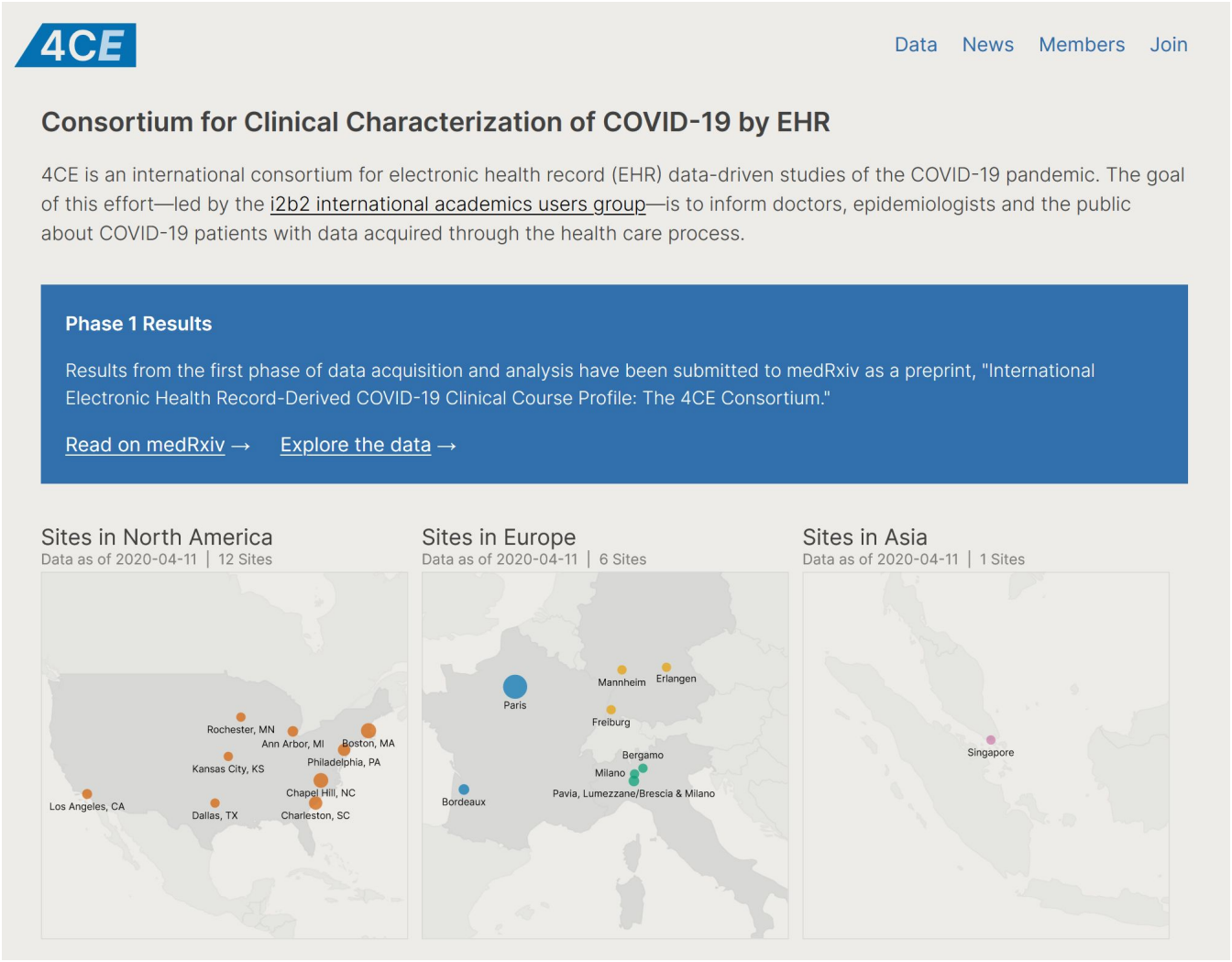
- Animal Behavior and Cognition
- Biochemistry
- Bioengineering
- Bioinformatics
- Biophysics
- Cancer Biology
- Cell Biology
- Clinical Trials
- Developmental Biology
- Ecology
- Epidemiology
- Evolutionary Biology
- Genetics
- Genomics
- Immunology
- Microbiology
- Molecular Biology
- Neuroscience

Konkretne problemy

Electronic Health Record (EHR)

Źródło: https://ehrintelligence.com/news/how-professional-services-ensure-successful-ehr-adoption

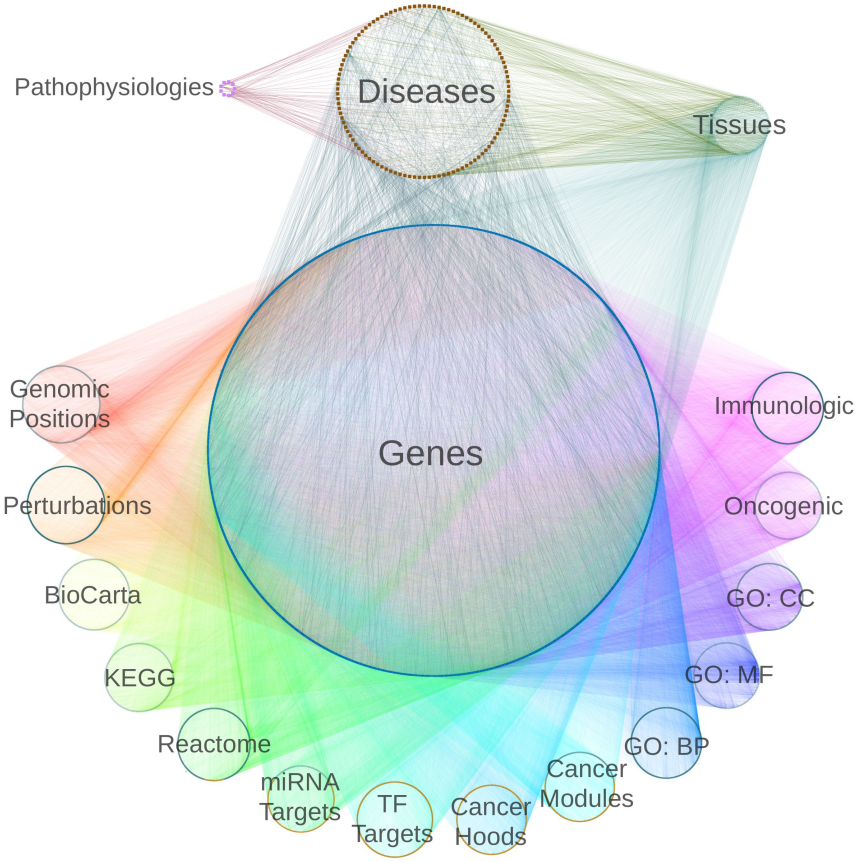
EHR - Covid 19



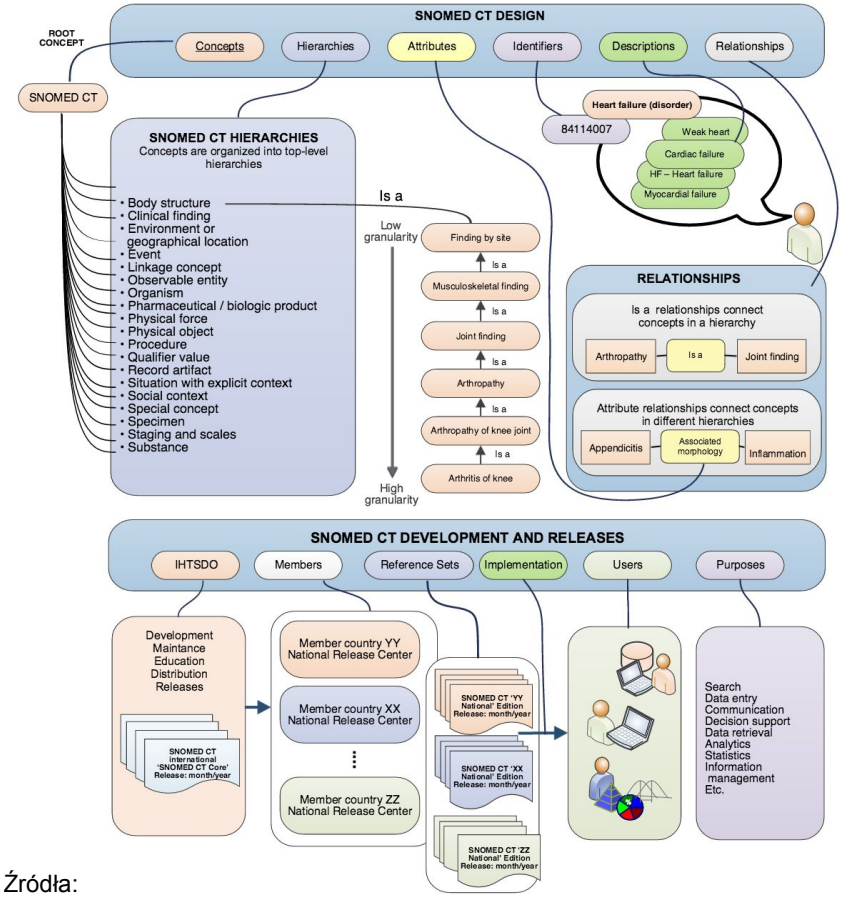
Źródło: https://covidclinical.net/

Graph-based knowledge

het.io



SNOMED CT



Źródła:
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004259
https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics

Niektóre projekty

- Biclustering / EBIC
- Benchmarking ML: fairness, objectiveness
- Triclustering - analiza szeregów czasowych
- AutoML / TPOT
- Feature selection / scikit-rebate
- Bioinformatics
  - big data analytics
  - drug repurposing
  - drug adverse effects
  - knowledge extraction
  - gene expression analysis
  - survival analysis (Kaplan-Maier estimator)
  - EHR - Covid-19

Manifesto

- ✓ UNDERSTAND YOUR DATA!!!
- ✓ What is your goal? Model interpretability or performance?
- ✓ Understand the limitations of your model.
- ✓ Do not trust the results.
- ✓ Better doesn’t mean significantly better.
- ✓ Make your code reproducible.
- ✓ Hyper-parameter tuning is very important.
- ✓ Stratify, cross-validate.
- ✓ Automate, automate (if only possible) ...
- ✓ Deep learning is not a solution to every problem.
- ✓ Publish your source code open source.
- ✓ *No free lunch theorem*
- ✓ *“All models are wrong, but some are useful”* (George Box)

Dziękuję za uwagę!

- ✓ Biclustering / Triclustering
- ✓ Machine learning
- ✓ Artificial Intelligence
- ✓ Benchmarking: fairness, objectiveness
- ✓ AutoML
- ✓ Feature selection
- ✓ Bioinformatics
  - big data
  - EHR
  - gene expression analysis
  - knowledge and association mining
  - drug repurposing

patryk.orzechowski@gmail.com
patrick@agh.edu.pl
patryk@upenn.edu

github.com/athril
linkedin.com/in/patryk.orzechowski
home.agh.edu.pl/patrick
Google Scholar

Perelman School of Medicine University of Pennsylvania

Institute for Biomedical Informatics

AGH

Katedra Automatyki i Robotyki