Hackathon CERN@TUe... A few insights from JADS!

Just to get you started ☺

Damian A. Tamburri, Ph.D.

TU/e Eindhoven University of Technology

Tilburg University

JADS Jheronimus Academy of Data Science

Damian A. Tamburri, Ph.D.

Tenure: Associate Professor, DataOps Systems Engineering at the TU/e and JADS

JADS lectures:

- Big Data Engineering
- Deep Learning
- Machine-Learning (pre-master)

Research interests: Data-Intensive Services DevOps/DataOps, Social Software Engineering, and Artificial-Intelligence Software.

Publication Records: Damian has published over 150+ papers in either top Journals or conferences in Software Engineering, Information Systems, and Services and AI Computing.

Perks: Damian is passionate about Dungeons&Dragons, video games, comics, plays the piano and is very sociable so ask any question you wish of him and he'll do the best to answer!

Favorite poison: Amuerte Gin with Fever-Tree Indian Tonic, so ask him about Gin&Tonics as well if you wish!

We do cool stuff that matters, with data

TU/e Eindhoven University of Technology

Tilburg University

JADS Jheronimus Academy of Data Science

Dr. Dario Di Nucci

Tenure: Assistant Professor, Machine-Learning and AI Software at UniTilburg and JADS

JADS lectures:
- Big Data Engineering
- Machine-Learning (pre-master)

Research interests: Software Maintenance and Evolution, Artificial-Intelligence Software.

Publication Records: Dario has published over 50+ papers in either top Journals or conferences in Software Engineering, Information Systems, and AI Computing.

Perks: Dario is passionate about …

Favorite poison: He'll drink anything you push in front of him… including GASOLINE!

TU/e Eindhoven University of Technology · Tilburg University · JADS Jheronimus Academy of Data Science

Dr. Gemma Catolino

Tenure: Post-Doc, Machine-Learning, AI Software biases, Organisational and Social Aspects of AI and Software Engineering at UniTilburg and JADS

JADS lectures:

- Machine-Learning (pre-master)

Research interests: Software Maintenance and Evolution, Artificial-Intelligence Software.

Publication Records: Gemma has published over 30+ papers in either top Journals or conferences in Software Engineering, Information Systems, and AI Computing.

Perks: Gemma is passionate about …

Favorite poison: She is the heart and soul of applied AI in our beloved JADE Lab… RESEARCH is her favourite poison! Naah, she likes Gin and (as the picture suggests) she also enjoys her Prosecco :D

TU/e Eindhoven University of Technology
Tilburg University
JADS Jheronimus Academy of Data Science

Ekhtiar Syed

Tenure: Adjunct Professor at TU/e and JADS, Product Manager at ASML, responsible for commercial and technical management of our data products

JADS lectures:

- Data Engineering (MSc)
- Data Engineering (PDEng)

Research interests: Big Data Engineering, Artificial-Intelligence Software, Applied AI.

Publication Records: Ekh has worked on any number of kernels and other AI-related products published in the past few years;

Perks: Ekhtiar just loveZ to participate into Keras challenges and tournaments and has won several prizes and badges already, too many to report here

Favorite poison: Ekh loves his wine, but also mixes cocktails (if I'm not mistaken) :)

JADS Jheronimus Academy of Data Science

Now let's get serious…

“ **Problem in a nutshell:**

- **Pattern Matching and Allocation Job;**
- **Unsupervised dataset;**
- **Multivocal features;**
- **Dishomogeneous feature space;**

”

JADS Jheronimus Academy of Data Science

IDEA 0, start simple…

“ **K-Means**

(perhaps we even throw in some black-box optimization and/or boosting)

“clustering method based on vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.”

Cit. Wikipedia

”

IDEA 0, start simple…

“ K-Means ”

(perhaps we even throw in some black-box optimization and/or boosting)

“clustering method based on vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.”

Cit. Wikipedia

Makes zero to very simplistic assumptions… applies to our case ☺

Advantages:

1. Relatively simple to implement;
2. Scales to large data sets;
3. Guarantees convergence;
4. Can warm-start the positions of centroids;
5. Easily adapts to new examples;

…

JADS Jheronimus Academy of Data Science

IDEA 0-bis, still simple, but slightly less…

“ **HDBScan**

(perhaps we even throw in some black-box optimization and/or boosting)

“hierarchical clustering density-based approach to clustering with non-parametric assumptions […] focuses on finding low-density regions”

Cit. Wikipedia

”

JADS Jheronimus Academy of Data Science

IDEA 0-bis, still simple, but slightly less…

“ **HDBScan**

(perhaps we even throw in some black-box optimization and/or boosting)

“hierarchical clustering density-based approach to clustering with non-parametric assumptions […] focuses on finding low-density regions”

Cit. Wikipedia

Most widely used at scale and for Data-Intensive computing… both are true in our case!

”

Advantages:

1. Widely used, lots of improvements;
2. Scales to large data sets;
3. very good at separating dense from sparse data;
4. robust to many outliers/noise;
…

JADS Jheronimus Academy of Data Science

IDEA 1

“ Gaussian-Mixture Modelling

(perhaps we even throw in some black-box optimization method e.g., Newton-Raphson)

“A **Gaussian mixture model** is a probabilistic **model** that assumes all the data points are generated from a **mixture** of a finite number of **Gaussian** distributions with unknown parameters.
“ Cit. Scikit

”

JADS Jheronimus Academy of Data Science

IDEA 1

“ Gaussian-Mixture Modelling

(perhaps we even throw in some black-box optimization method e.g., Newton-Raphson)

“A **Gaussian mixture model** is a probabilistic **model** that assumes all the data points are generated from a **mixture** of a finite number of **Gaussian** distributions with unknown parameters. “ Cit. Scikit

This should apply in our case as well!

”

Advantages:

1. Modelling with unknown subpopulation any data point belongs to;
2. Learn the subpopulations automatically, (e.g., often used for weak supervision);
3. Fuzzy classification of observations - can be combined with other methods;

JADS Jheronimus Academy of Data Science

IDEA 2

“Robust Random Cut Forest (RRCF)

(perhaps we even throw in some advanced bootstrapping methods)

“ensemble technique for detecting outliers. The idea is based on an isolation forest algorithm that uses an ensemble of trees.“ Cit. Figueirêdo et al. [1]

”

JADS Jheronimus Academy of Data Science

IDEA 2

“Robust Random Cut Forest (RRCF)”

(perhaps we even throw in some advanced bootstrapping methods)

“ensemble technique for detecting outliers. The idea is based on an isolation forest algorithm that uses an ensemble of trees.“ Cit. Figueirêdo et al. [1]

This addresses several characteristics of our dataset

Advantages:

1. Different dimensions are treated independently;
2. Designed to run in streaming data;

JADS Jheronimus Academy of Data Science


And Now… CRAZY
IDEAS!

IDEA 3

“ Deep Convolutional Neural Networks Augmented with Unsupervised Feature Learning

“ Why not using a convolutional approach (typically used very successfully in whether pattern prediction and forecasting [2]) but augmented with unsupervised feature learning [3] such that either weak supervision or latent variable discovery can kick in… “

”

JADS Jheronimus Academy of Data Science

IDEA 3

“ Deep Convolutional Neural Networks Augmented with Unsupervised Feature Learning

“ Why not using a convolutional approach (typically used very successfully in whether pattern prediction and forecasting [2]) but augmented with unsupervised feature learning [3] such that either weak supervision or latent variable discovery can kick in… “

Note to self: THIS WOULD BE AWESOME IF IT WORKS!

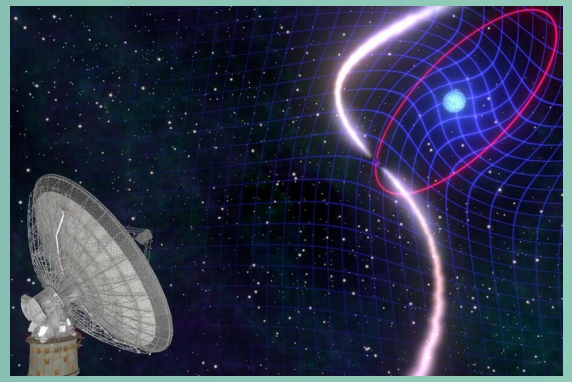
”

JADS Jheronimus Academy of Data Science

IDEA 4

“ Space-Time Convolutional Neural Networks

“ A recurrent neural network architecture specifically designed for forecasting time series of spatial processes, i.e., series of observations sharing temporal and spatial dependencies [4]“



”

JADS Jheronimus Academy of Data Science

IDEA 4

“ Space-Time Convolutional Neural Networks

“ Why not using a recurrent neural network architecture specifically designed for forecasting time series of spatial processes, i.e., series of observations sharing temporal and spatial dependencies?“

Cit. Eric Postma

”

JADS Jheronimus Academy of Data Science

IDEA 4

“ Space-Time Convolutional Neural Networks

“ Why not using a recurrent neural network architecture specifically designed for forecasting time series of spatial processes, i.e., series of observations sharing temporal and spatial dependencies?“

Cit. Eric Postma

Many of the assumptions behind the usage of this technique apply in this scenario (or do they!?)

”

JADS Jheronimus Academy of Data Science

Last but not least…

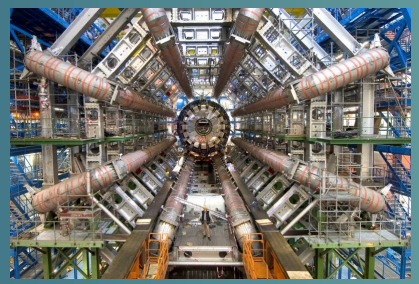
IDEA 5

– more like, best fit idea that goes first (if you want to solve the problem fast) or last (if you want to inspire more outta-da-box thinking) ☺

Go to the roots: CLUstering of Energy approach

"density-based clustering algorithm, optimized for high-occupancy scenarios, where the number of clusters is much larger than the average number of hits in a cluster. The algorithm uses a grid spatial index for fast querying of neighbors and its timing scales linearly with the number of hits within the range considered. "

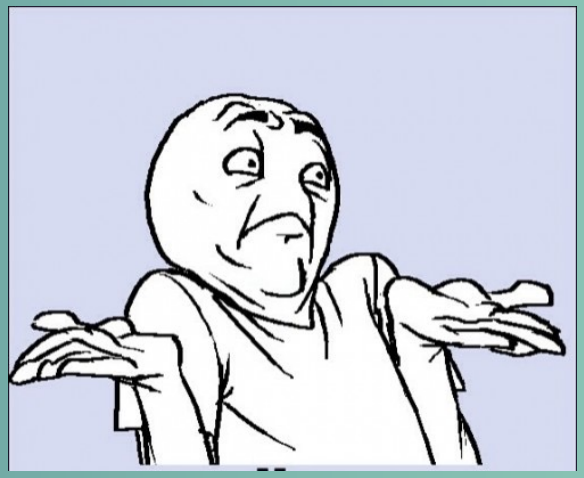
Cit. Rovere et al. [5]

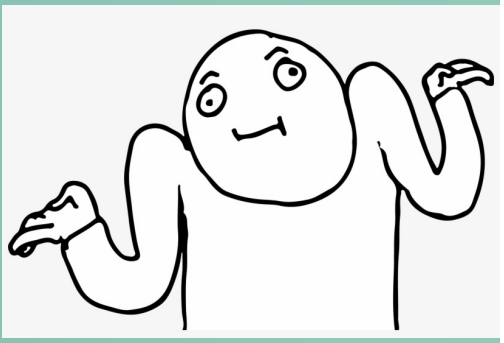


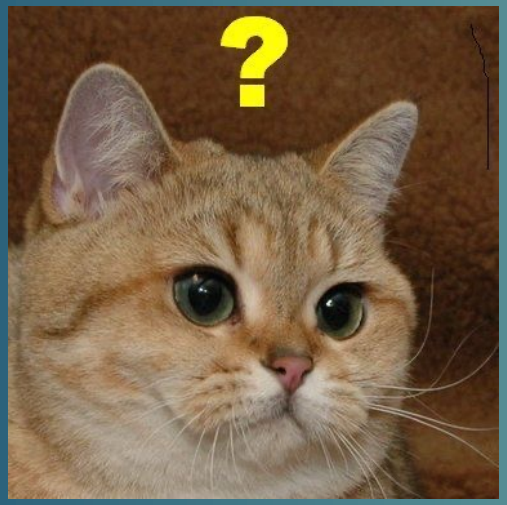
Ok… That's it!

"

•Questions?







"

JADS Jheronimus Academy of Data Science

Bibliography

[1] “Multivariate Real Time Series Data Using Six Unsupervised Machine Learning Algorithms” Ilan Figueirêdo, Lílian Lefol Nani Guarieiro and Erick Giovani Sperandio Nascimento, 2020.

[2] Liu, Y., Racah, E., Prabhat, Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., & Collins, W. (2016). Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets. ArXiv, abs/1605.01156.

[3] K. Nguyen, C. Fookes and S. Sridharan, "Improving deep convolutional neural networks with unsupervised feature learning," 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 2270-2274.

[4] Ziat, A., Delasalles, E., Denoyer, L. & Gallinari, P. (2017). Spatio-Temporal Neural Networks for Space-Time Series Forecasting and Relations Discovery.. In V. Raghavan, S. Aluru, G. Karypis, L. Miele & X. Wu (eds.), ICDM (p./pp. 705-714), : IEEE Computer Society. ISBN: 978-1-5386-3835-4

[5] Rovere Marco, Chen Ziheng, Di Pilato Antonio, Pantaleo Felice, Seez Chris, «A Fast Parallel Clustering Algorithm for High Granularity Calorimeters in High-Energy Physics» Frontiers in Big Data, 3, 2020, URL=https://www.frontiersin.org/article/10.3389/fdata.2020.591315

JADS Jheronimus Academy of Data Science