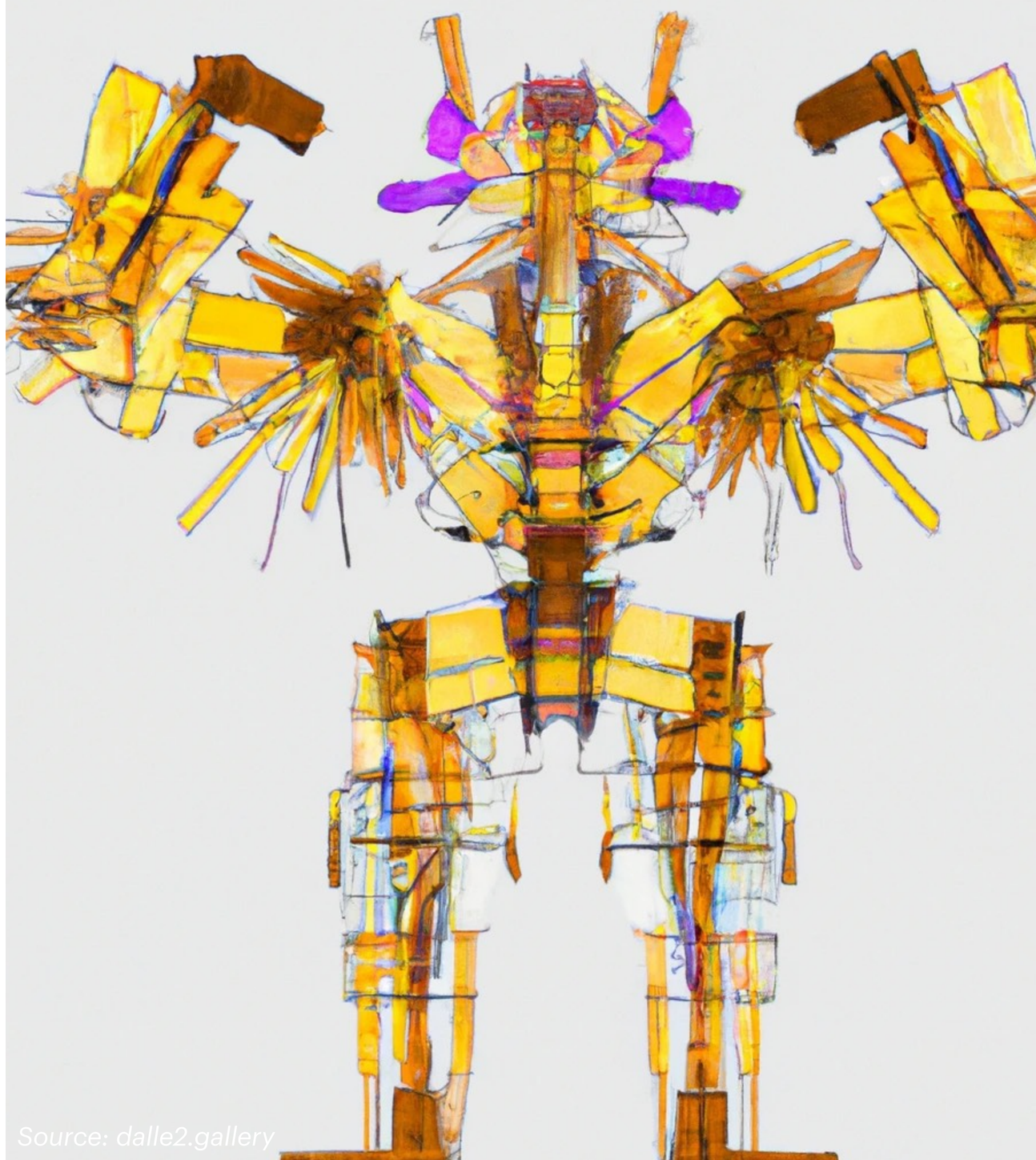
Unlocking Data Insights - Introduction to Data-Centric AI

New life to your data

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Future Artificial Intelligence Research
UniBa UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
KDDE

Data-Centric AI: transforming raw
Executive summary

- Traditional encoding methods
- Advanced encoding methods
- Nice Tools



Source: dalle2.gallery

Traditional encoding methods

001

100

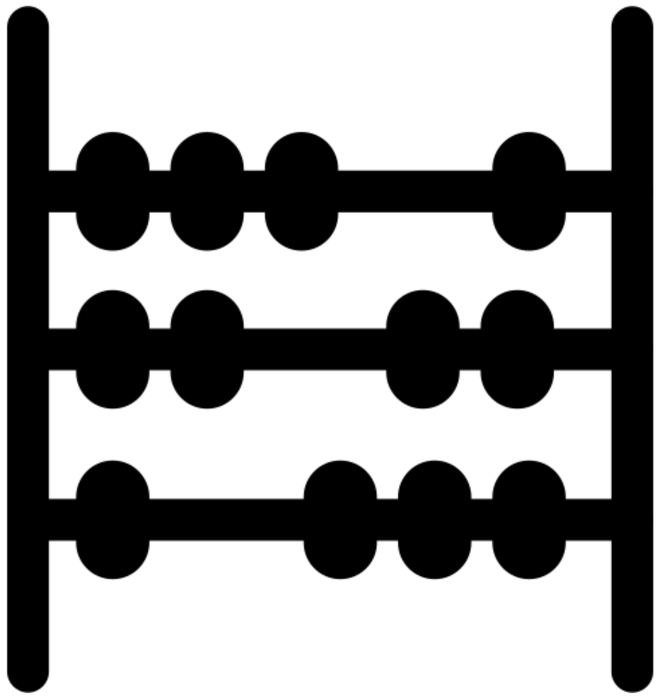
010

One-hot (Weiss et al., 2015): given a variable containing n different values, the variable is transformed into an array where each unique value is represented as a binary vector with the $i-th$ position set to one and the rest set to zero.

Traditional encoding methods

Limits?

Traditional encoding methods



CountVectorizer (count2vec) (Weiss et al., 2015): given a collection of categorical documents, this method produces a matrix of token occurrences, where each line in the matrix represents a document and each column a token. The size of the vector space depends on the n unique values in the vector space.

Traditional encoding methods

Limits?

Traditional encoding methods

TF-IDF (Luhn, 1958): the term frequency (TF) captures the frequency of a particular token w.r.t. to a given document, whereas the inverse document frequency (IDF) measures how common the token is in the corpus.

TF = n. of times the term appears in the document/total number of terms in the document

IDF = log(n. of the document in the corpus/number of the documents in the corpus contain term)

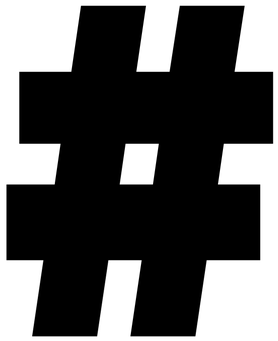
TF-IDF = TF * IDF

Traditional encoding methods

Limits?

Traditional encoding methods

HashVectorizer (hash2vec) (Weiss et al., 2015): it does the same as count2vec. However, instead of storing tokens, it directly maps each token to a column position in the matrix of occurrences. It is mainly useful for large datasets and unlike one-hot and count2vec, which have the same dimensionality as the vocabulary length, this method has the flexibility to hash tokens in any dimensionality.

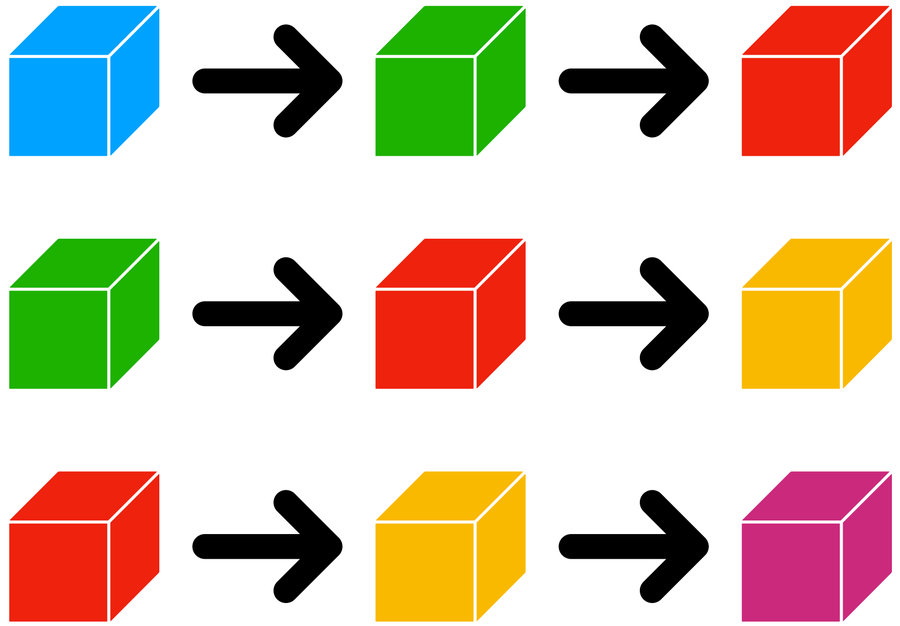


Traditional encoding methods

Limits?

Traditional encoding methods

N-grams (Gasparetto et al., 2022): this method represents a given sequence of elements through sub-sequences of n items. Thus, considering a sequence $\mathbf{s} = \{s1, \ldots, si\}$, the n-grams representation of these sequences is given by $n - grams = \{(s1, \ldots, sn), (s2, \ldots, sn{+}1), \ldots (si{-}n, \ldots, si)\}$.



Traditional encoding methods

Limits?

Advance encoding methods

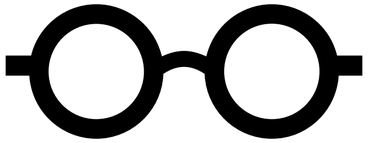
Source: dalle2.gallery

Encoding methods are responsible for transforming complex information into a representative feature space, i.e., mapping data from one space to another*

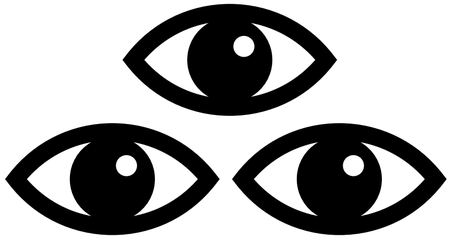
*Gabriel M. Tavares, Rafael S. Oyamada, Sylvio Barbon, Paolo Ceravolo, Trace encoding in process mining: A survey and benchmarking, Engineering Applications of Artificial Intelligence, Volume 126, Part D, 2023

Why we need the new data form?

To unlock the predictive power of Neural Networks

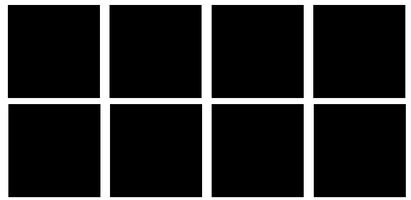


Discover possible pattern into to the data



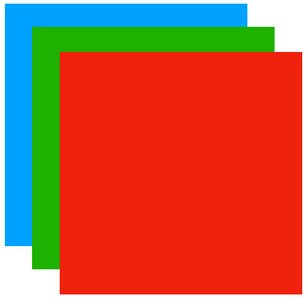
Combine different data sources

Advanced encoding methods



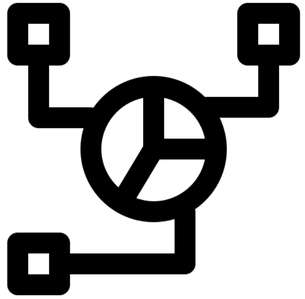
Contextual Emedding

Word2Vec
LLMs: Bert ...



Image

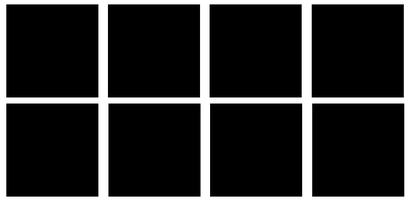
Grey Scale Image
RGB Image



Hybrid approach

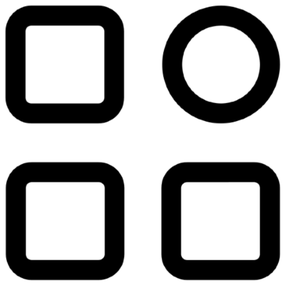
Combine encoding methods

Contextual embedding



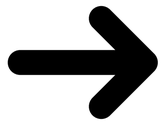
1. Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.
2. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34: 18932–18943
3. Liu, Guang, Jie Yang, and Ledell Wu. "PTab: Using the Pre-trained Language Model for Modeling Tabular Data." *arXiv preprint arXiv:2209.08060* (2022)

TabTransformer



Categorical value in a feature can mean different things, depending on the context

e.g: **fox** in the forest
fox news

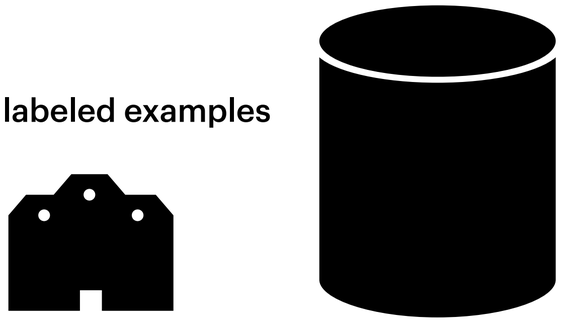


TabTransformer use that information in tabular data and represent the same value with differently, if the context is different

Why adopt TabTransformer?

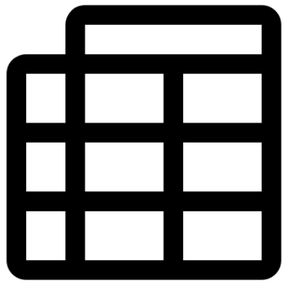
A lot of unlabeled examples

Few labeled examples



Pre-training procedure on unlabeled examples + fine-tuning

Pre-training:
Masked Language Modeling (MLM)
Replace Token Detection (RTD)

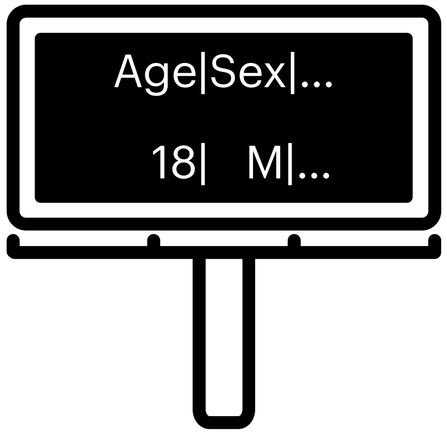


The majority of datasets in production are tabular

Keras implementations

https://keras.io/examples/structured_data/tabtransformer/

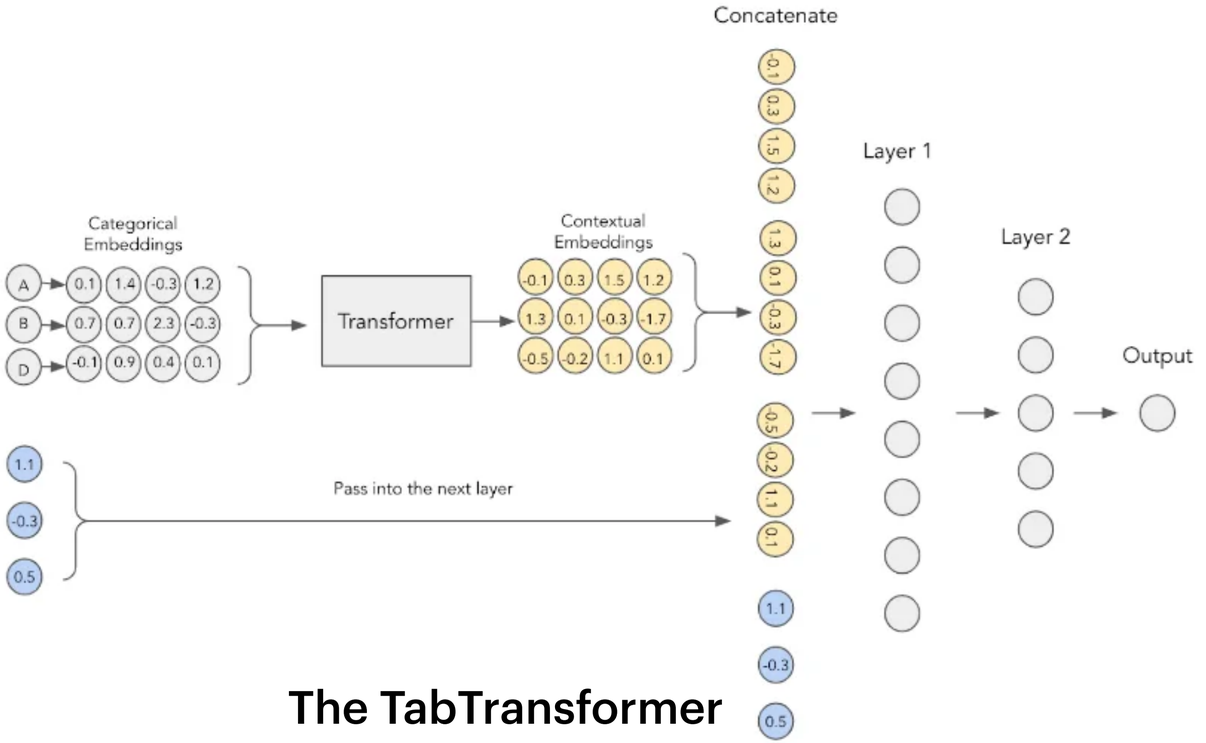
TabTransformer



Tabular Data
e.g. Pandas Dataframe

Categorical Feature

Numerical Feature



The TabTransformer architecture
*Credits: reddit**

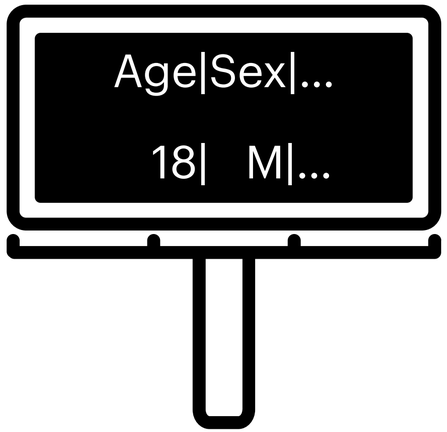
*https://www.reddit.com/media?url=https://preview.redd.it/project-improving-deep-learning-for-tabular-data-with-v0-mk28f629uxw91.png?width=1916&format=png&auto=webp&s=cfb25443c2235131ced58f8936cd97054ceabaf6

FT-Transformer

An extended version of
TabTransformer

FT-Transformer model transforms all features (categorical and numerical) to embeddings and applies a stack of Transformer layers to the embeddings

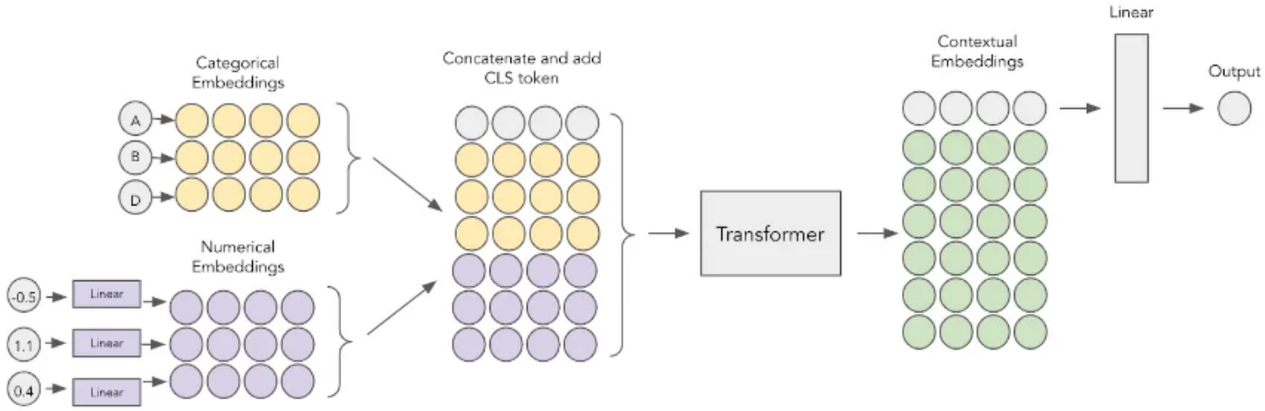
FT-Transformer



Tabular Data
e.g. Pandas Dataframe

Categorical Feature

Numerical Feature



The FT-Transformer architecture
*Credits: reddit**

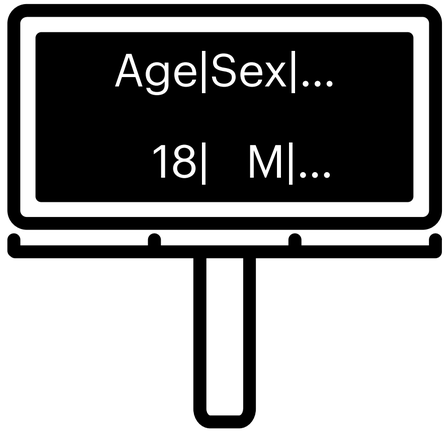
*https://www.reddit.com/media?url=https://preview.redd.it/project-improving-deep-learning-for-tabular-data-with-v0-mk28f629uxw91.png?width=1916&format=png&auto=webp&s=cfb25443c2235131ced58f8936cd97054ceabaf6

Tabular data as text

Thanks to the availability of several pre-trained LLMs we can

Transform our original dataset through Sentence embedding

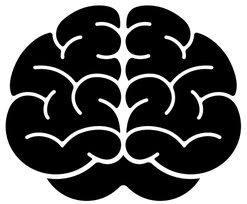
Tabular data as text [3]



Tabular Data
e.g. Pandas Dataframe

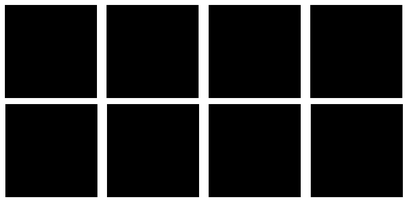
Job: Advocate,
Sex: M, ...

Transform column in
text



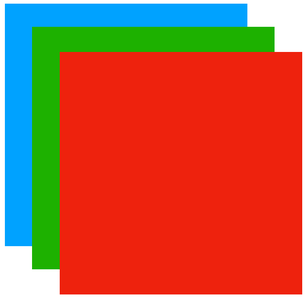
Generate Embedding
with pre-trained model

Advanced encoding methods



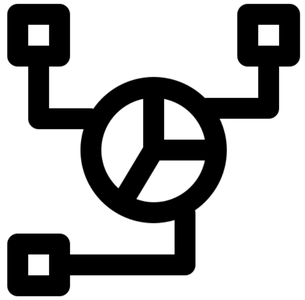
Contextual Emedding

Word2Vec
LLMs: Bert ...



Image

Grey Scale Image
RGB Image



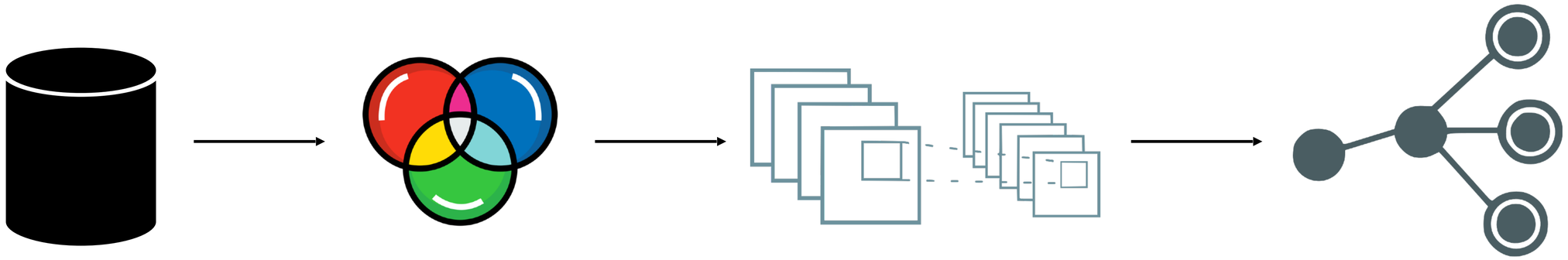
Hybrid approach

Combine encoding methods

Data like image

1. T. Kim, S. C. Suh, H. Kim, J. Kim and J. Kim, "An Encoding Technique for CNN-based Network Anomaly Detection," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle
2. Predictive Process Mining meets Computer Vision, Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba BPM Forum 2020
3. A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture", *Sci. Rep.*, vol. 9, no. 1, pp. 11399, Dec. 2019
4. ORANGE: Outcome-Oriented Predictive Process Monitoring Based on Image Encoding and CNNs Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, Donato Malerba, and Giuseppe Modugno, October 2020 IEEE Access Volume 8

Data like RGB image [1,2]



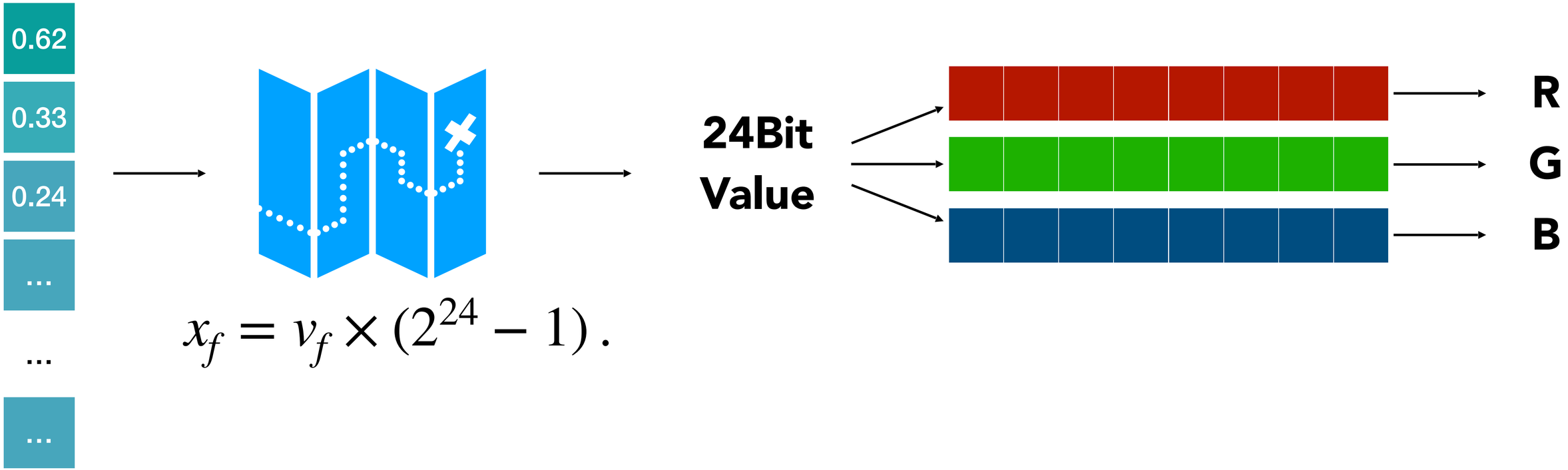
Tabular Data

RGB Encoding

Predictive model (CNN, ViT, etc.)

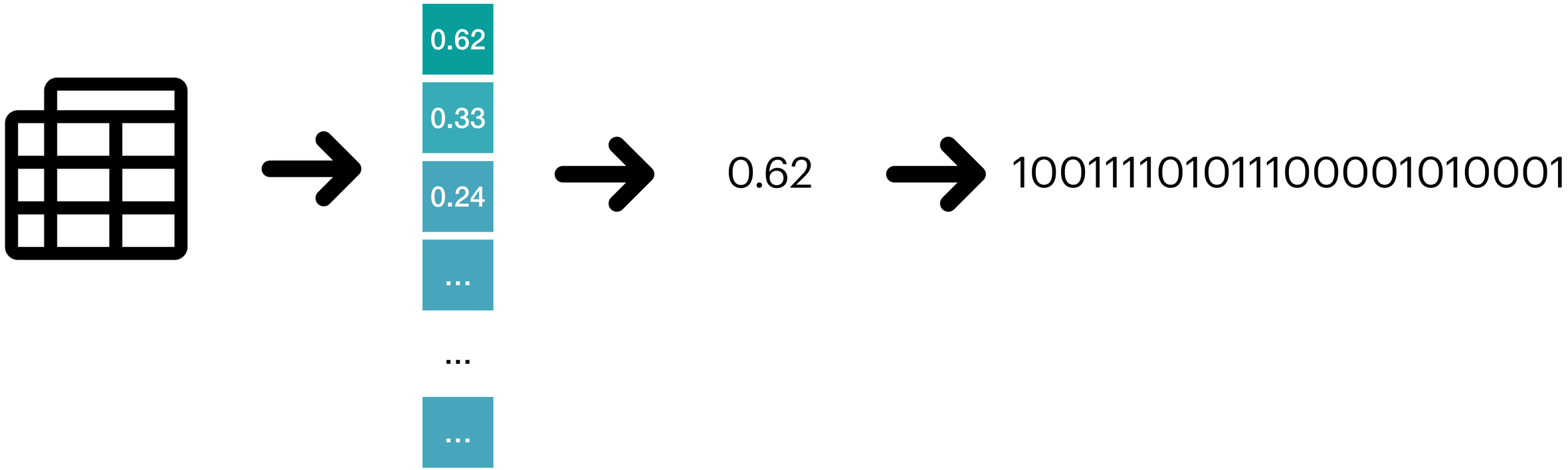
Prediction

Data like RGB image [1,2]



$$x_f = v_f \times (2^{24} - 1)\,.$$

Data like RGB image [1,2]



Tabular Data
e.g. Pandas Dataframe

Min-max
Normalization (0,1)

Transform into
24 bit value

Decimal to
binary

Data like RGB image [1,2]

10011110 10111000 01010001

Split into 3 groups (8 bit)

158 184 81

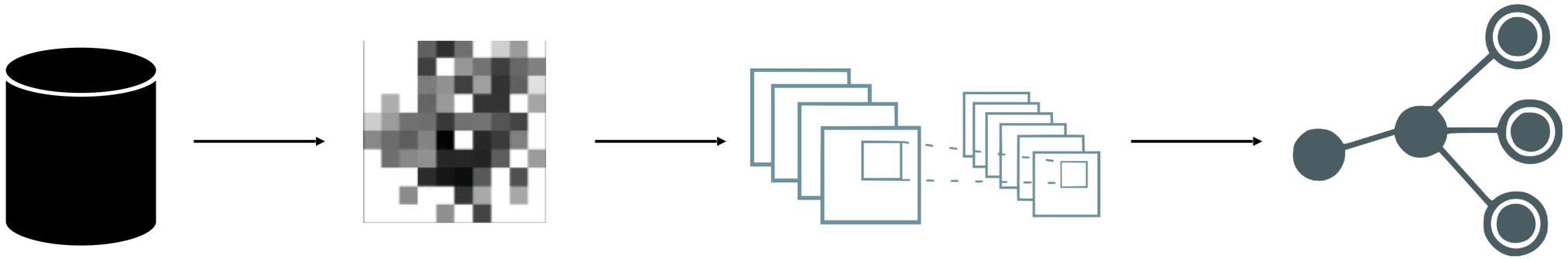
RGB Pixel value

Limits

Sequential layout of features

We are excluding possible relationships between features

DeepInsight encoding [3,4]



Tabular Data

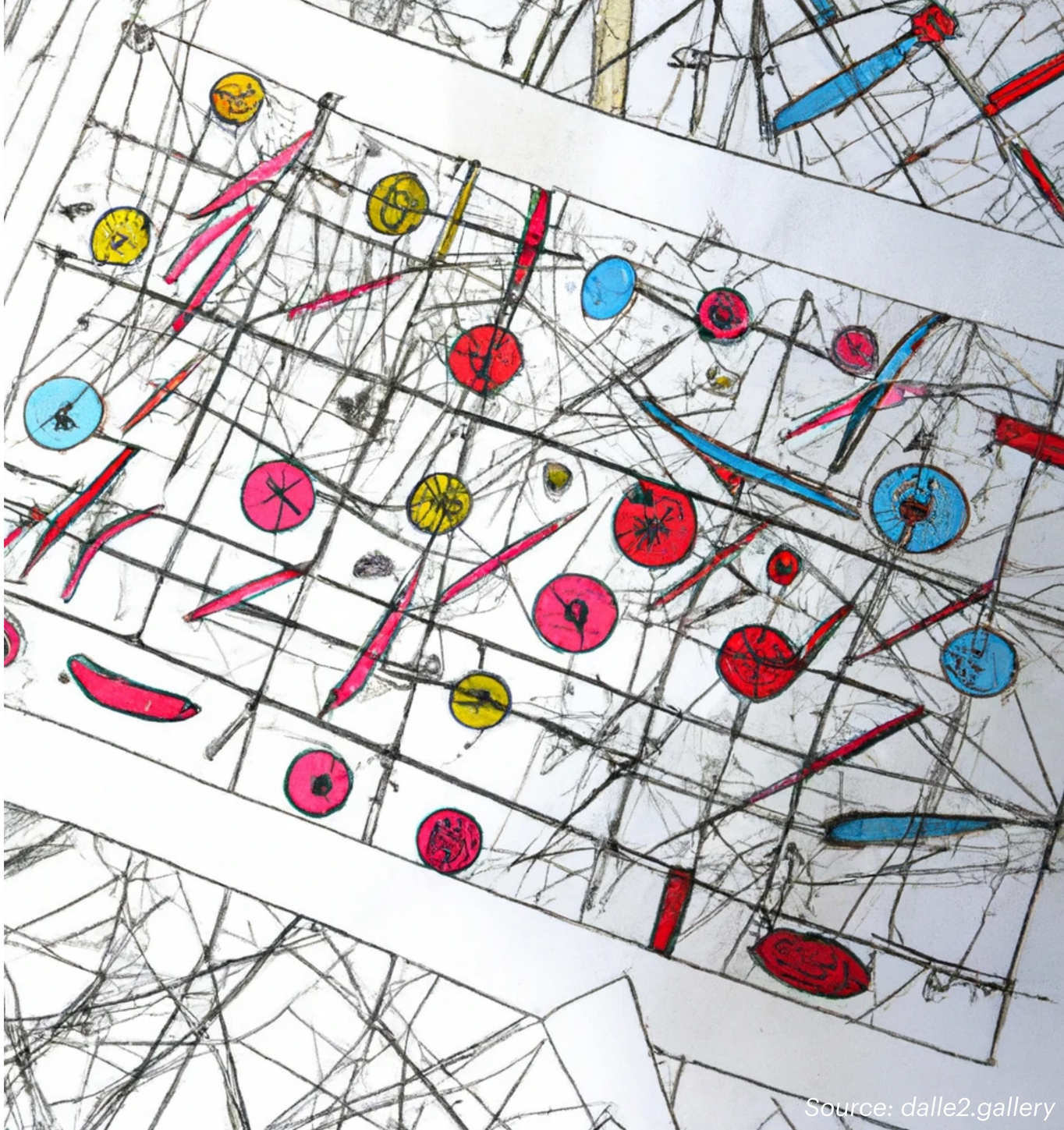
Grey scale Encoding

Predictive model (CNN, ViT, etc.)

Prediction

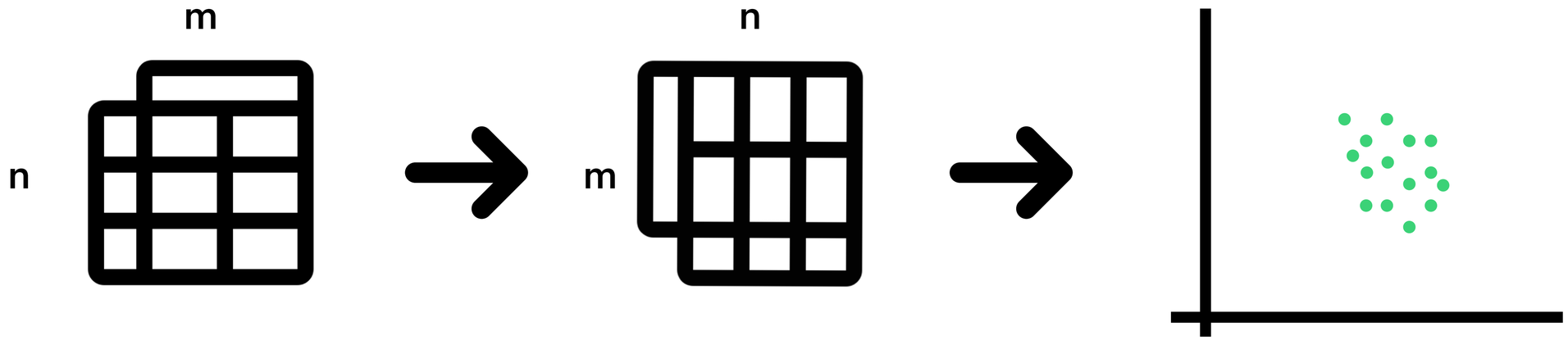
DeepInsight encoding

In an attempt to capture possible spatial relationships between features, the one-to-one association between features and pixel frames is done according to the theory introduced in [3]



Source: dalle2.gallery

DeepInsight encoding

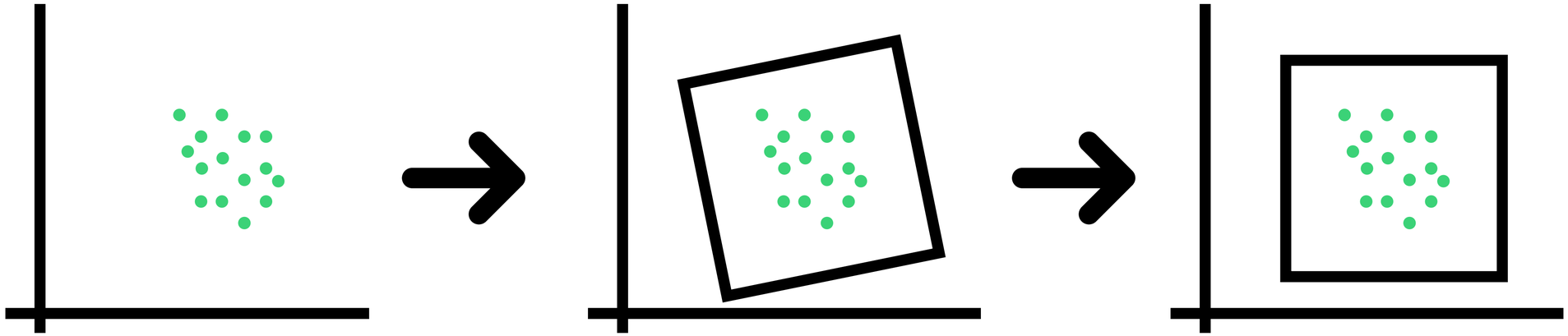


Tabular Data
e.g. Pandas Dataframe

Transpose

Dimensionality reduction
technique

DeepInsight encoding

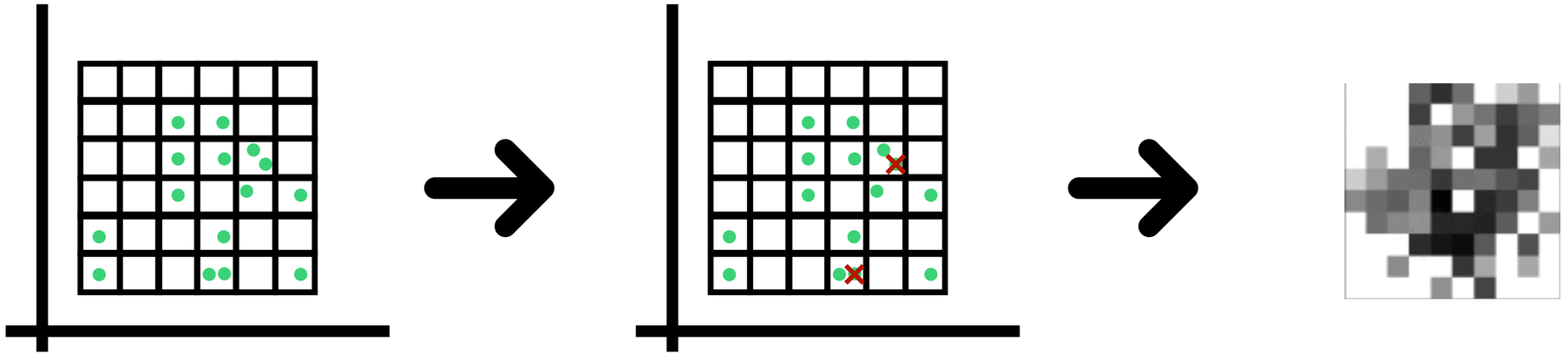


Dimensionality reduction technique

Convex Hull

Rotation

DeepInsight encoding

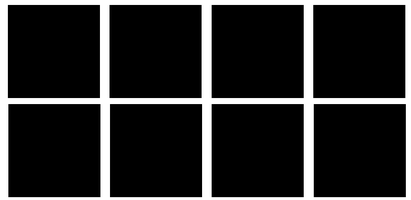


Grid definition

Detect possible collisions

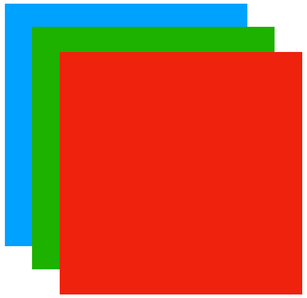
Grayscale image

Advanced encoding methods



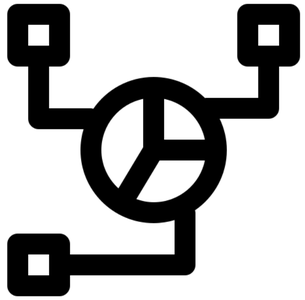
Contextual Emedding

Word2Vec
LLMs: Bert ...



Image

Grey Scale Image
RGB Image



Hybrid approach

Combine encoding methods

Hybrid Approach

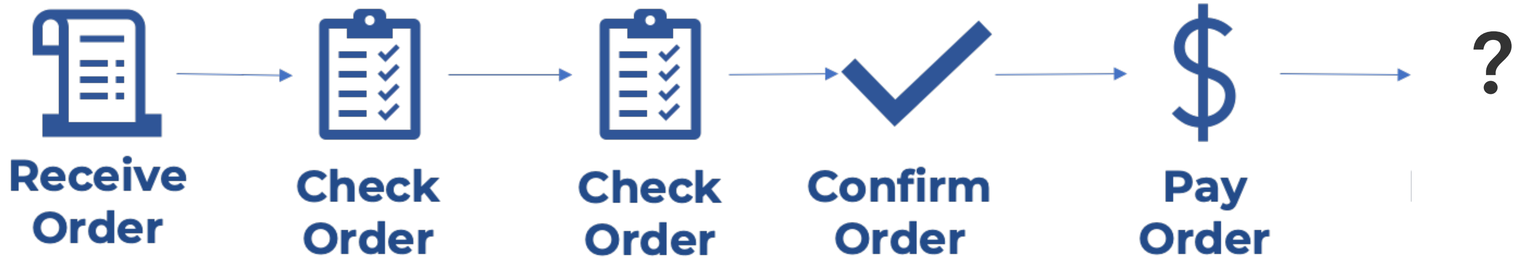
1. V. Pasquadibisceglie, A. Appice, G. Castellano and D. Malerba, "JARVIS: Joining Adversarial Training With Vision Transformers in Next-Activity Prediction," in IEEE Transactions on Services Computing

Combine different encoding methods to obtain new one

Hybrid Approach

Problem

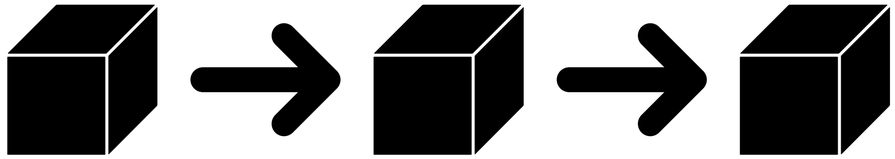
Develop a novel predictive process monitoring approach to solve the next-activity problem



Hybrid Approach

Existing solutions

If we consider the nature of the problem



→

The solution is to apply suitable encodings to handle sequences

Hybrid Approach

Existing solutions

In fact, most approaches in the literature adopt LSTM/Transformer neural networks

Considering different perspectives in an event log

Activity

Resource

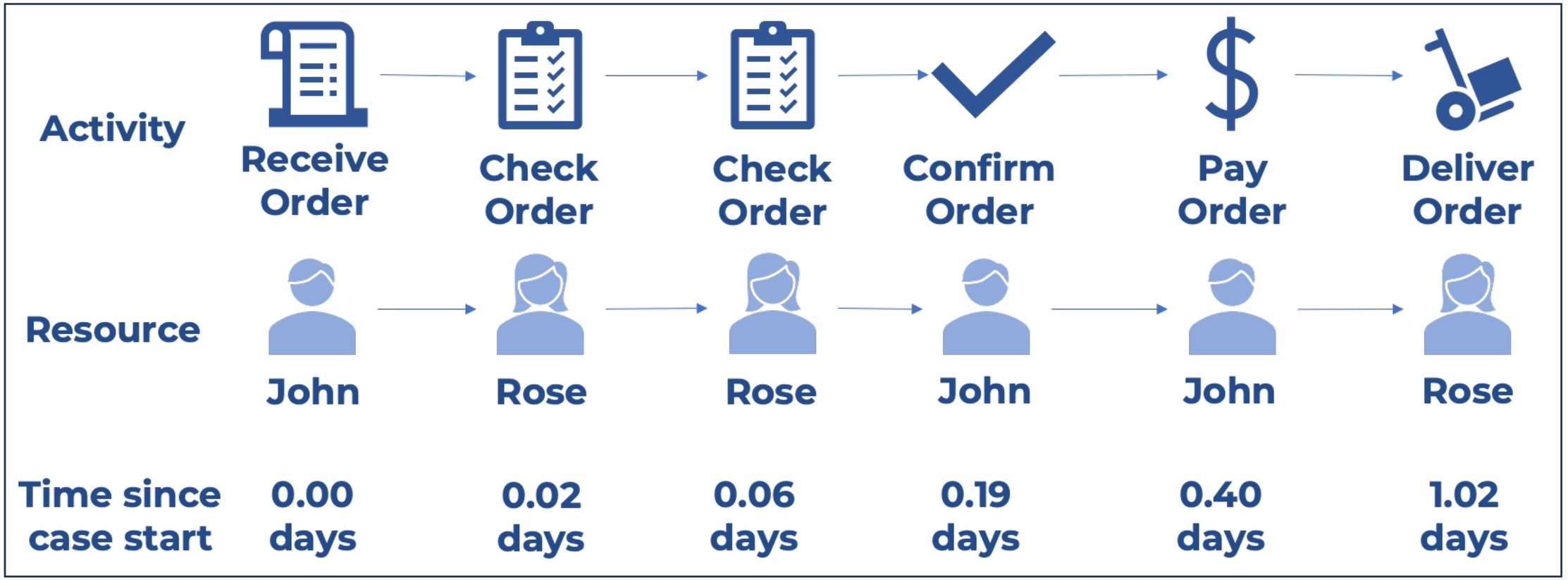
Timestamp

...

Hybrid Approach

Event log: An example

Event attribute

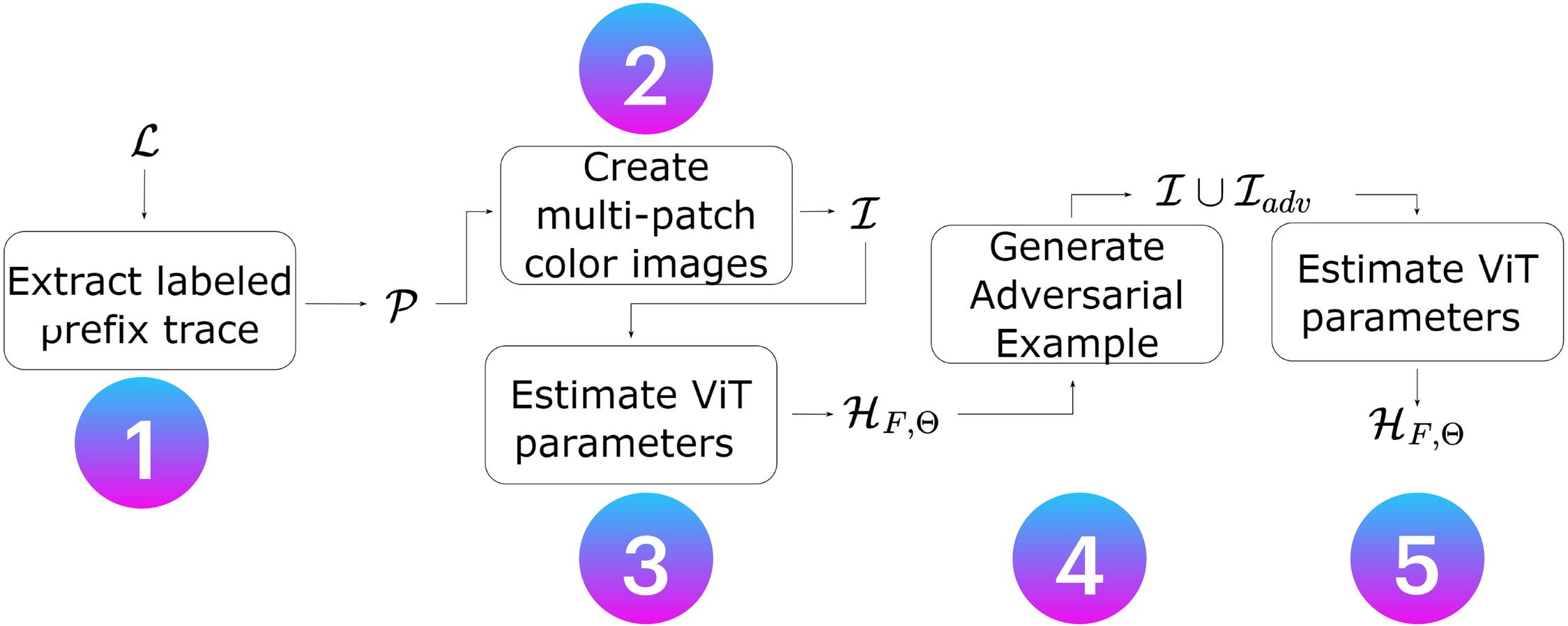


Trace attribute

Price	$45.00	$45.00	$45.00	$45.00	$45.00	$45.00

Hybrid Approach

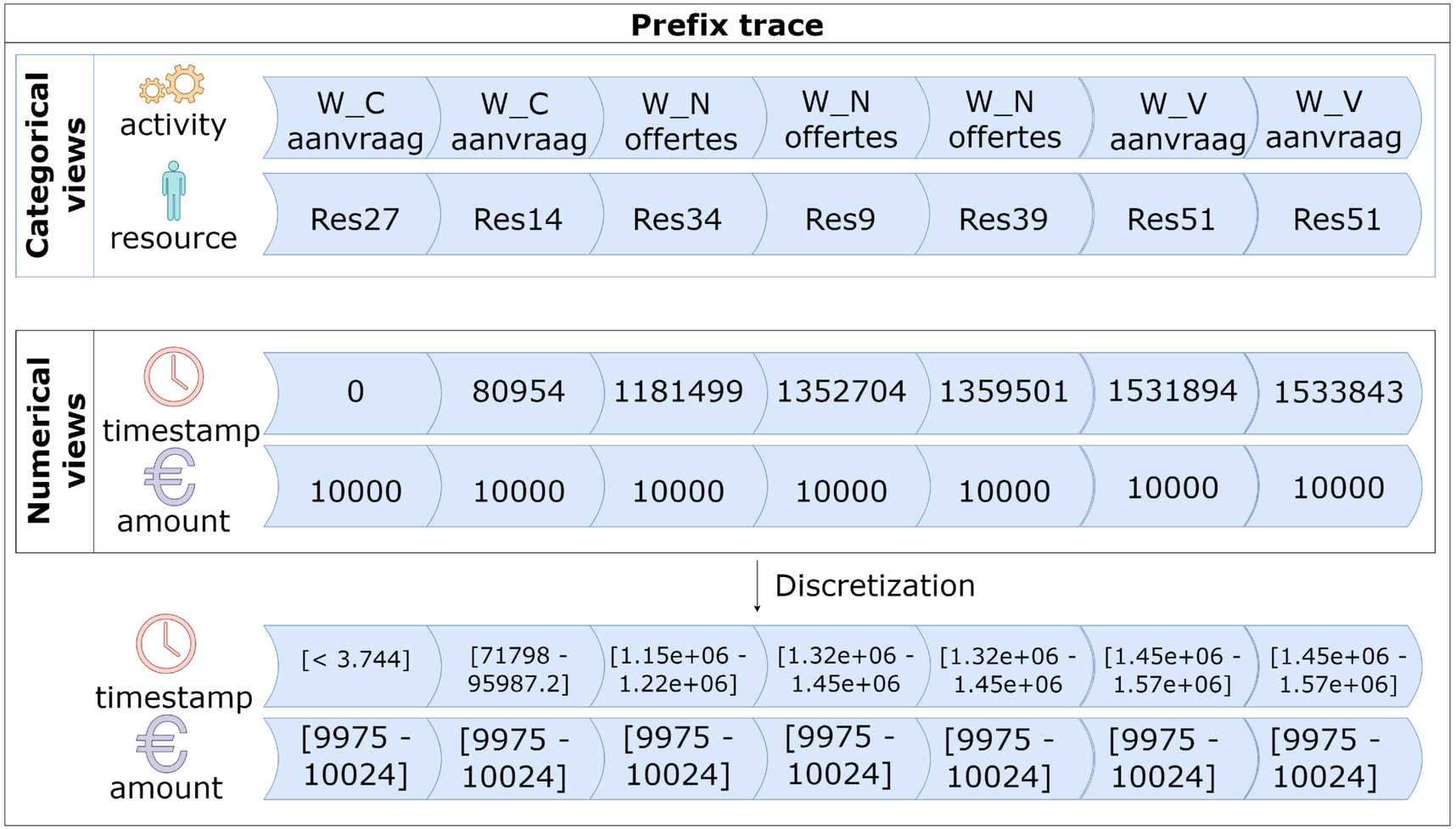
Proposed Approach: JARVIS



Hybrid Approach in action

Extract labeled prefix trace

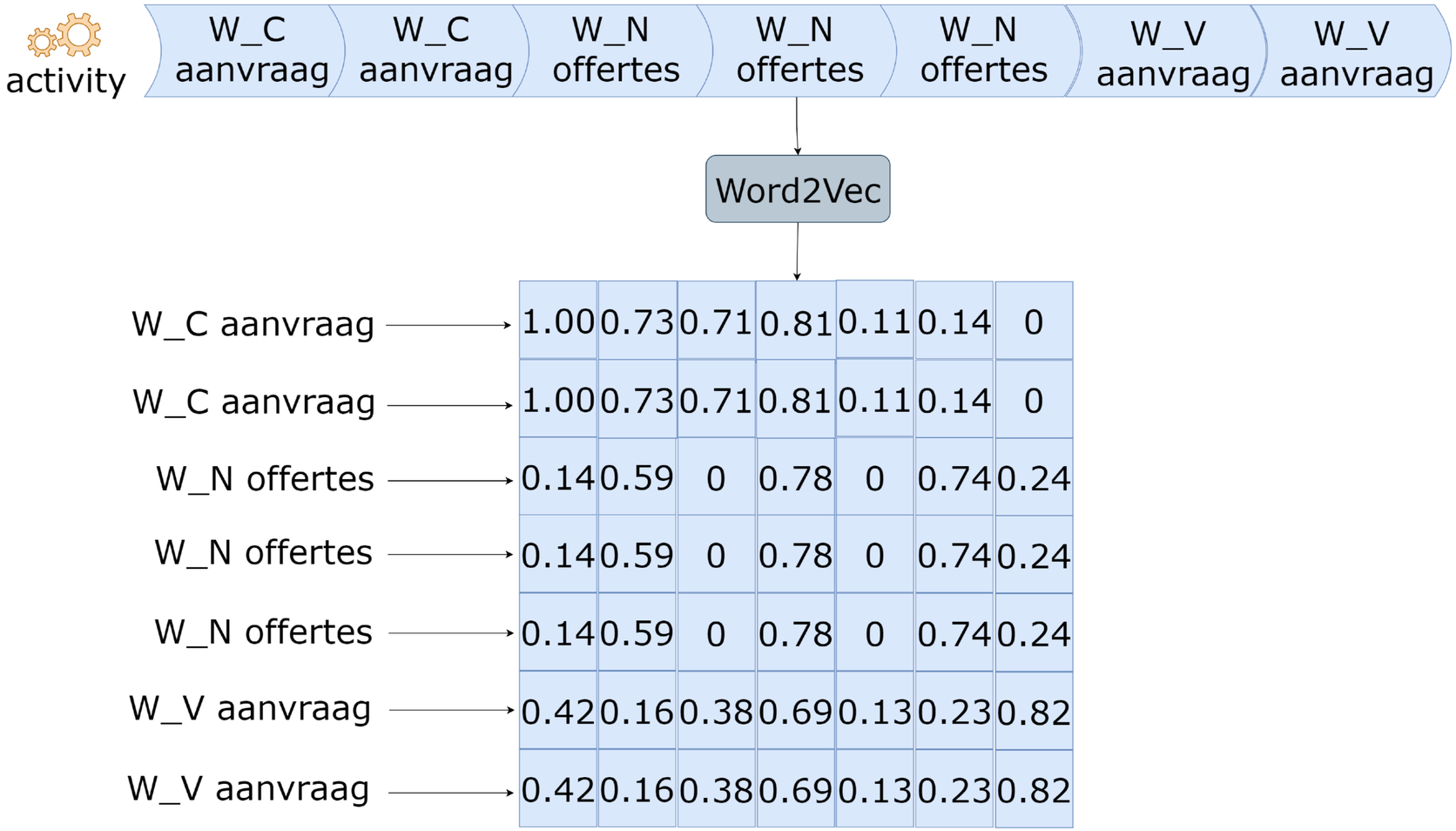
1



Hybrid Approach in action

Word2Vec embedding

1



Hybrid Approach in action

Word2Vec embedding to Imagery color patch

2

1.00	0.73	0.71	0.81	0.11	0.14	0
1.00	0.73	0.71	0.81	0.11	0.14	0
0.14	0.59	0	0.78	0	0.74	0.24
0.14	0.59	0	0.78	0	0.74	0.24
0.14	0.59	0	0.78	0	0.74	0.24
0.42	0.16	0.38	0.69	0.13	0.23	0.82
0.42	0.16	0.38	0.69	0.13	0.23	0.82

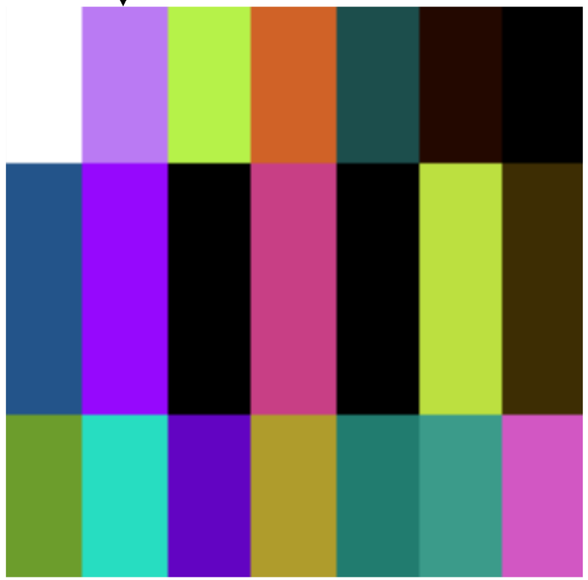
$$pixel8bit = 0.73 * (2^{24} - 1)$$

DecimalToBinary

[**10111010**, **11100001**, **01000110**]

BinaryToDecimal

[**R:186**, **G:122**, **B:243**]

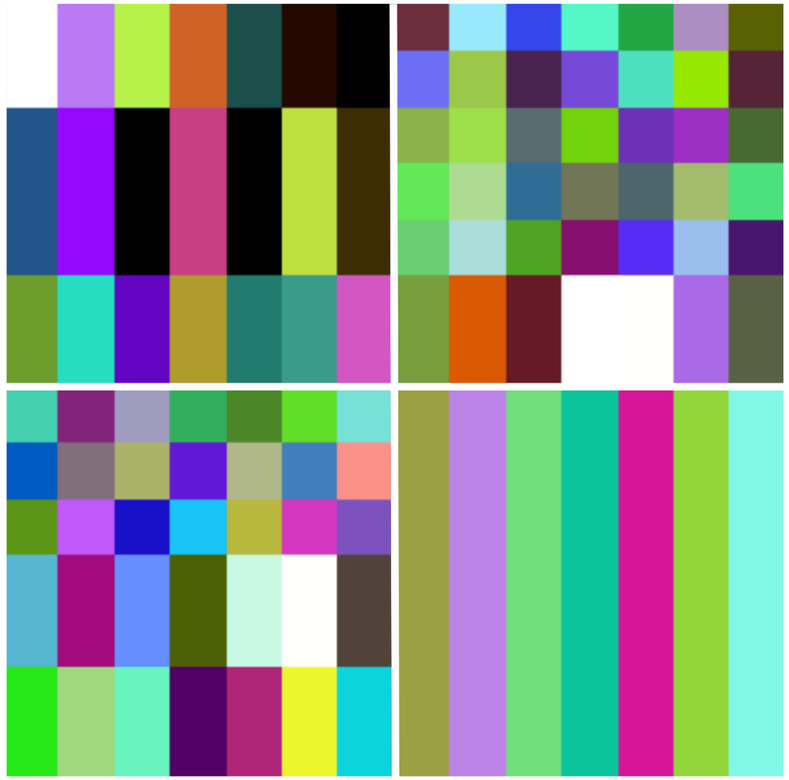


Hybrid Approach in action

Multi-patch color image

2

activity



resource

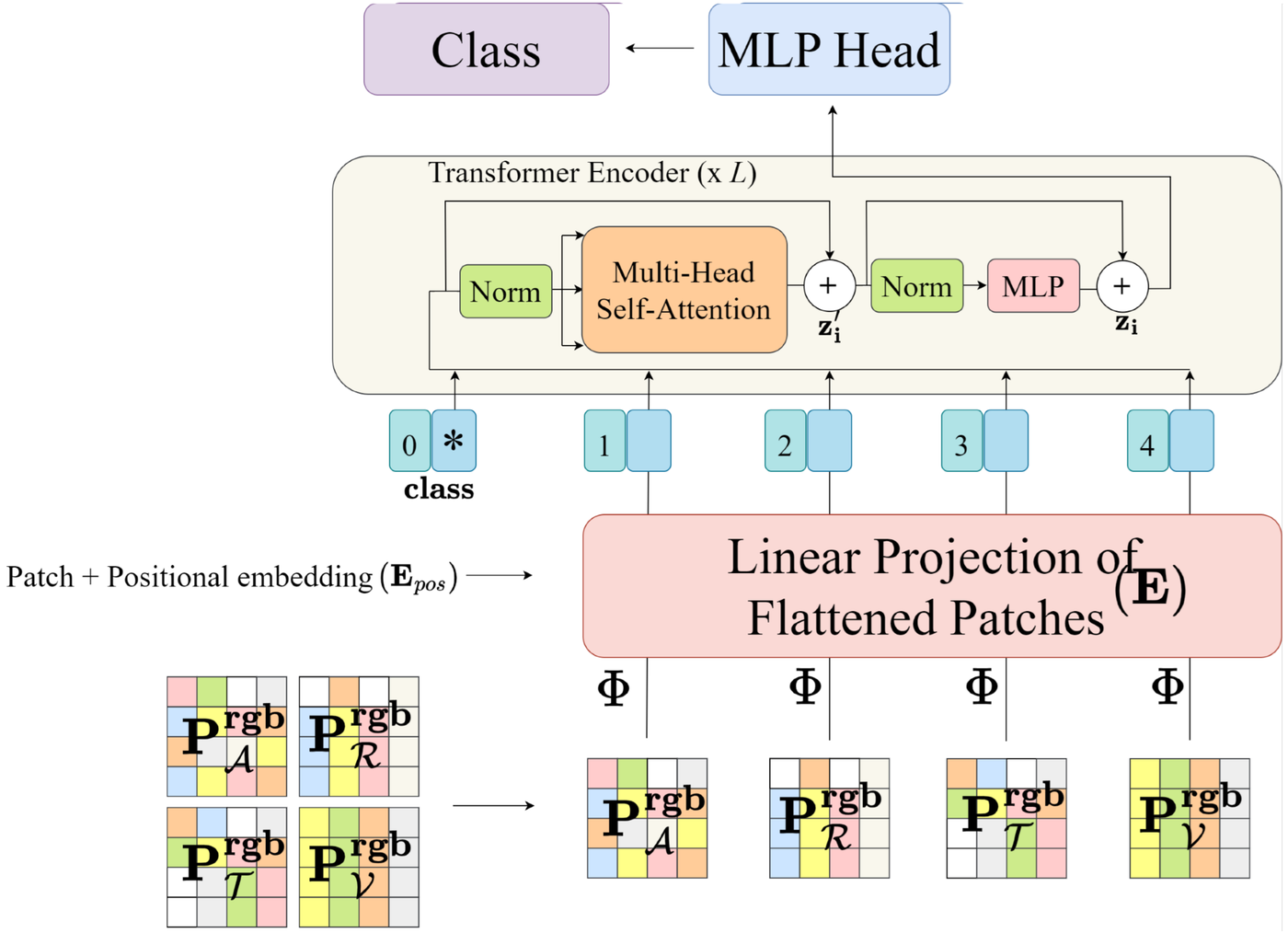
timestamp

amount

Hybrid Approach in action

Estimate ViT parameters

3



Hybrid Approach in action

Generate Adversarial Example

4

FGSM

PGD

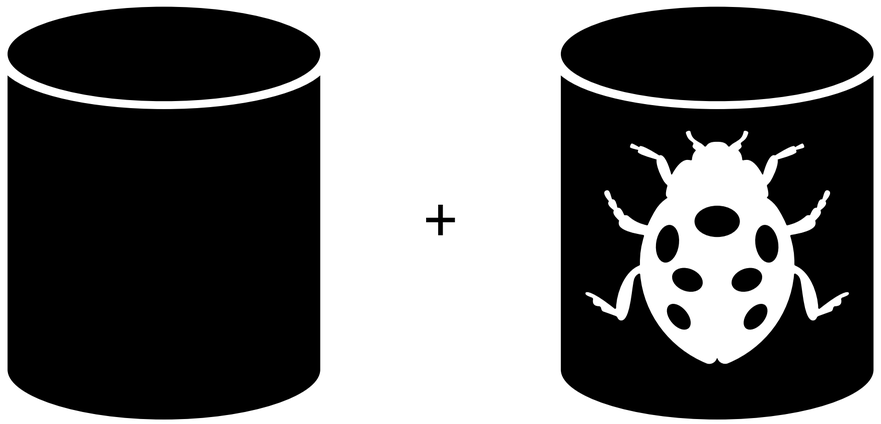
DeepFool

Hybrid Approach in action

Estimate the ViT parameters

5

Adversarial training

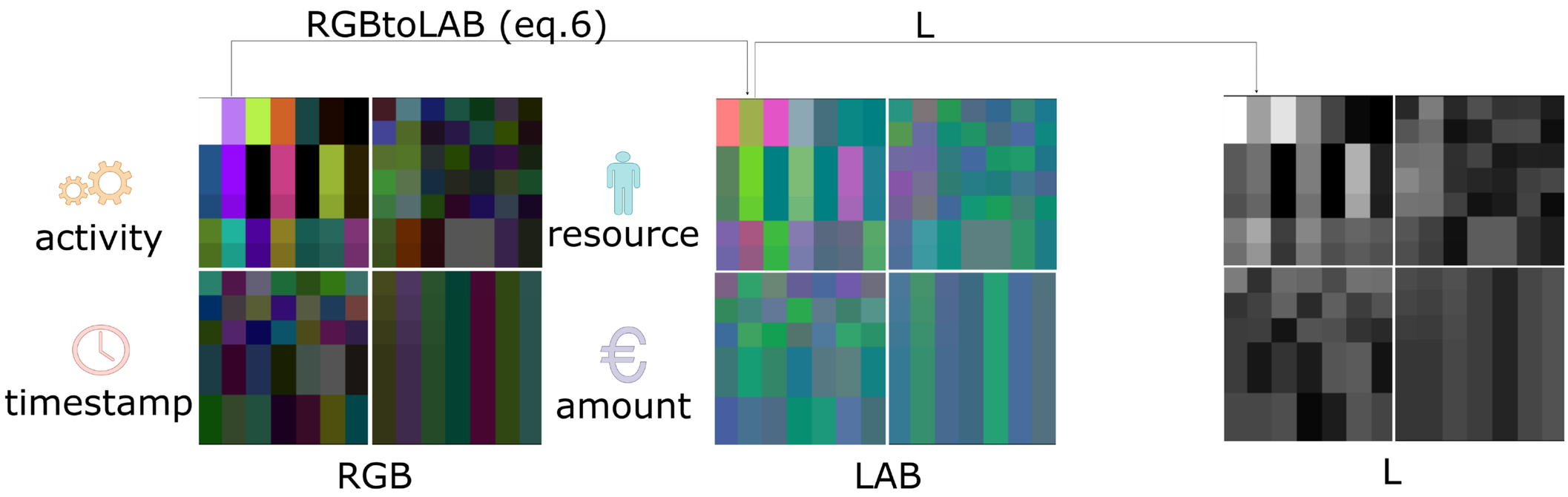


Original data

Adversarial data

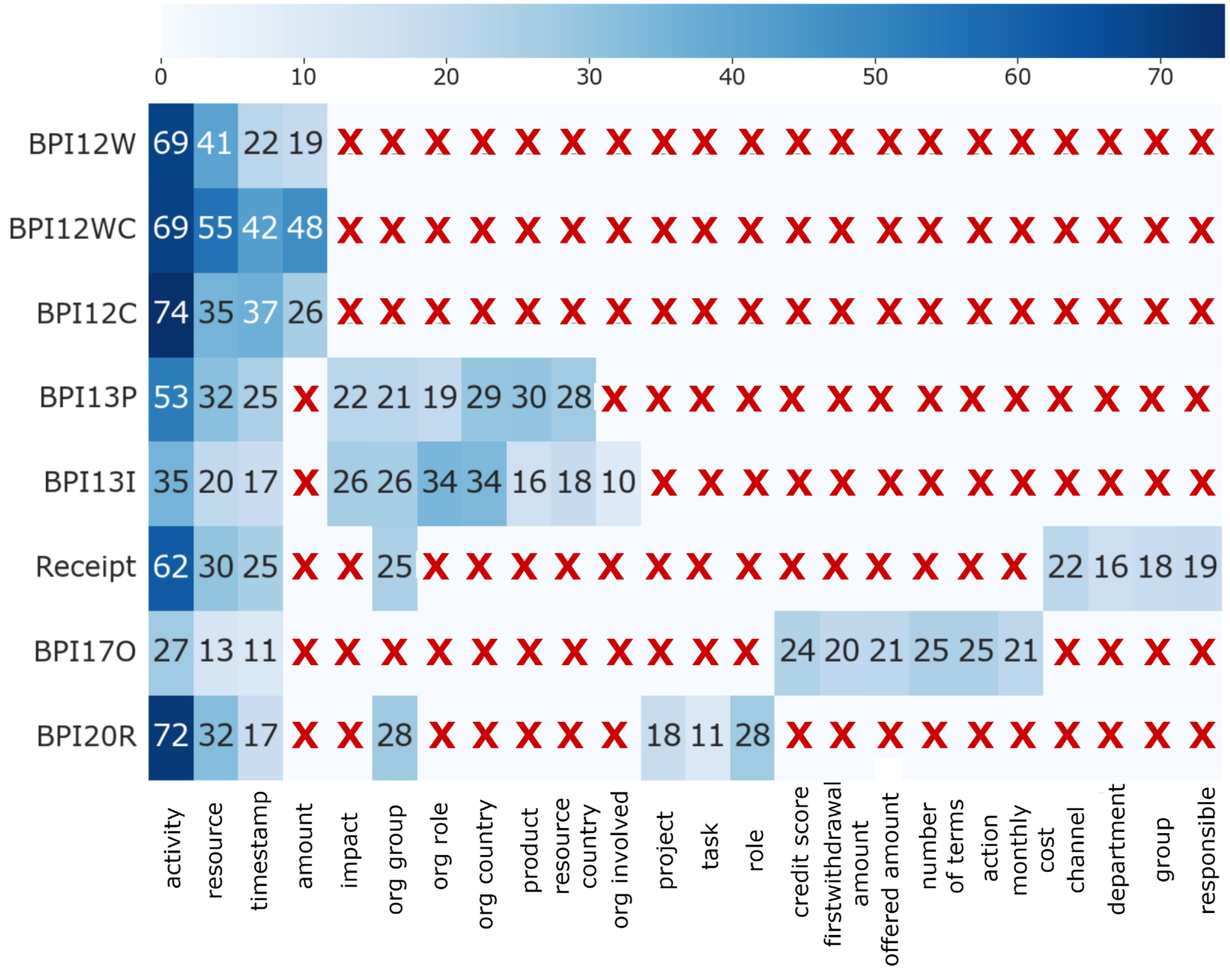
Hybrid Approach in action

Map of attention



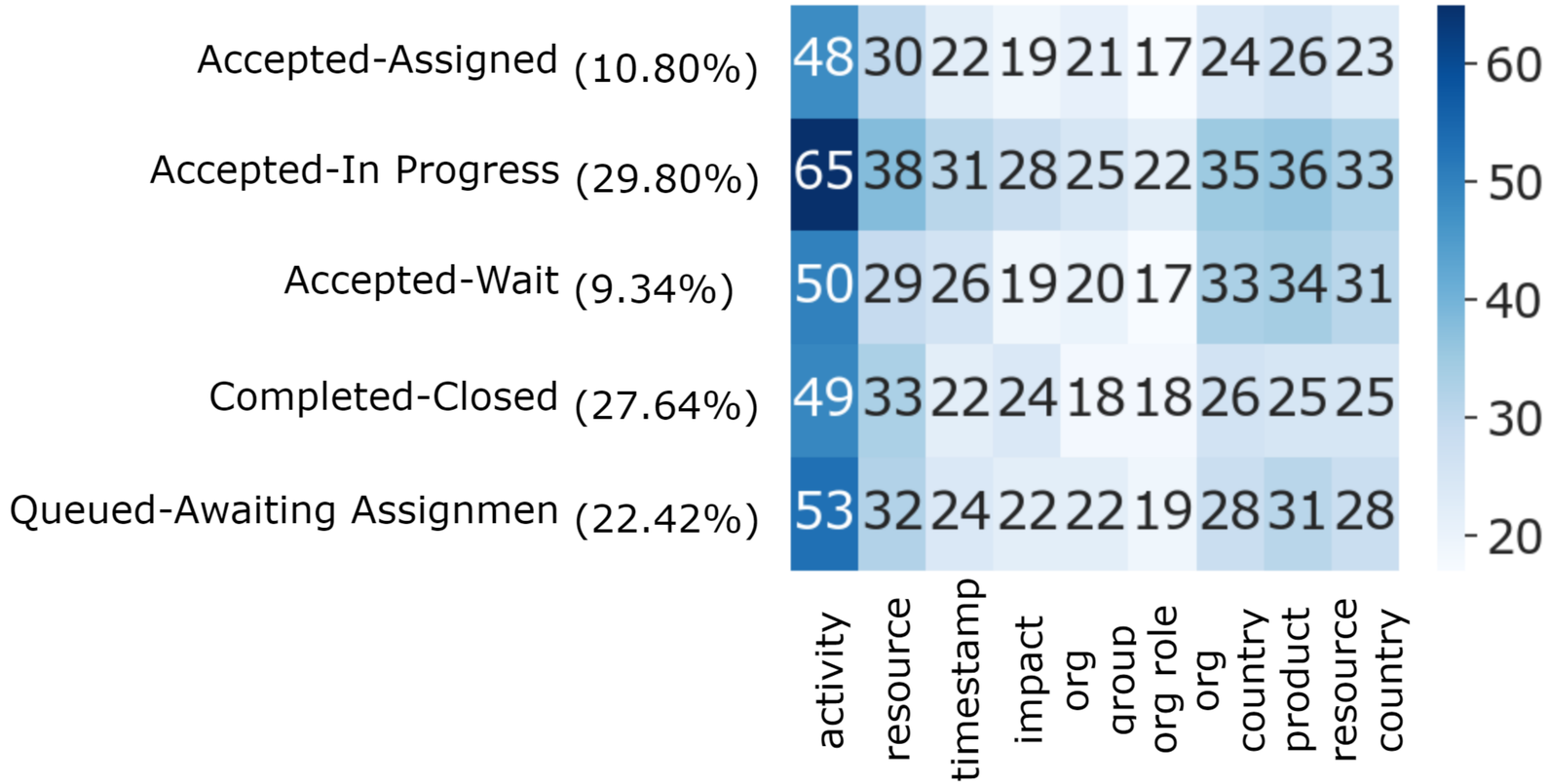
Hybrid Approach in action

Global explanation



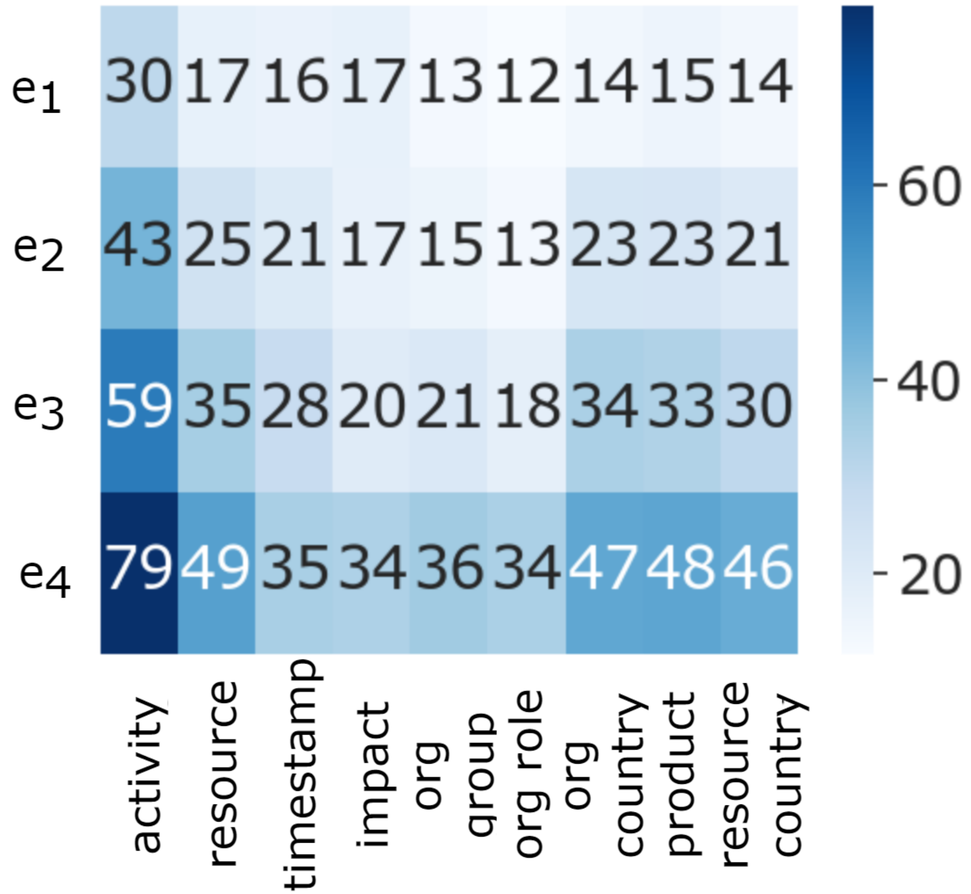
Hybrid Approach in action

Global explanation - Label analysis



Hybrid Approach in action

Global explanation - Event analysis



Hybrid Approach in action

Local explanation

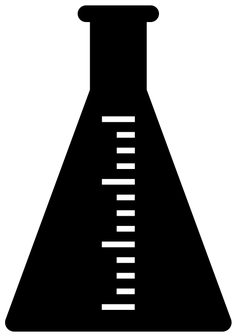
activity	Accepted in Progress	Accepted in Progress	Accepted Wait	Accepted in Progress
	108.0	108.0	188.0	76.0
org group	Org line C	Org line C	Org line C	Org line C
	72.25	112.25	111.25	105.25
product	Prod834	Prod834	Prod834	Prod834
	88.5	95.75	99.5	99.5

Hybrid Approach in action

Accuracy analysis

Eventlog	FScore						GMean					
	JARVIS	[3]	[4]	[9]	[20]	[23]	JARVIS	[3]	[4]	[9]	[20]	[23]
BPI12W	0.667	**0.737**	0.692	0.673	0.673	0.661	0.820	**0.847**	0.828	0.792	0.819	0.825
BPI12WC	**0.705**	0.685	0.661	0.675	0.645	0.668	**0.812**	0.798	0.778	0.792	0.780	0.787
BPI12C	0.644	**0.654**	0.642	0.638	0.643	0.624	0.786	**0.792**	0.782	0.785	0.781	0.781
BPI13P	**0.414**	0.320	0.336	0.408	0.228	0.405	**0.595**	0.533	0.546	0.594	0.472	0.593
BPI13I	0.387	0.405	0.295	**0.407**	0.363	0.380	0.615	**0.626**	0.534	**0.626**	0.594	0.603
Receipt	**0.525**	0.455	0.409	0.471	0.302	0.383	**0.733**	0.676	0.646	0.702	0.563	0.620
BPI17O	**0.720**	0.714	0.705	0.691	0.718	0.712	**0.846**	0.833	0.830	0.815	0.835	0.831
BPI20R	**0.491**	0.450	0.483	0.455	0.432	0.481	**0.699**	0.660	0.691	0.664	0.643	0.683

[3],[4] LSTM-based [9],[20] Image-based [23] Transformers



Nice Tools

https://colab.research.google.com/drive/1o_4QRq2llVqoB2c2RtjHGCUJqGOkZKUe?usp=sharing