

# **5<sup>th</sup> International Conference on Econometrics and Statistics (EcoSta 2022)**

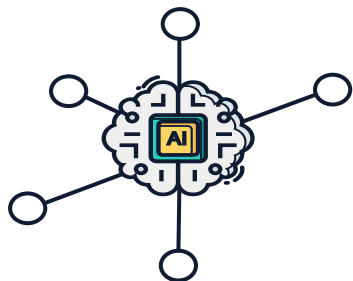
Ryukoku University, Kyoto, Japan, 4-6 June 2022

## **eXplainable AI for Finance**

Dr. Branka Hadji Misheva  
ZHAW, Switzerland

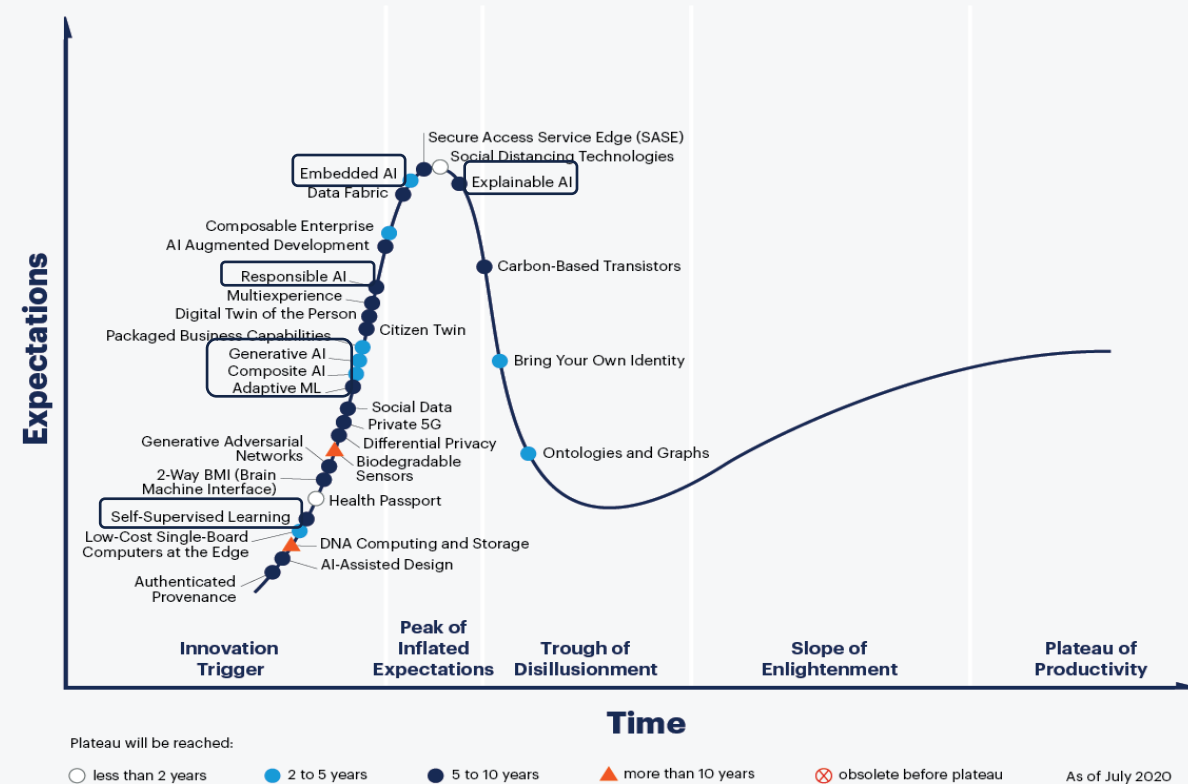
# The Need for XAI

Hype vs Real?



AI in finance?

## Hype Cycle for Emerging Technologies, 2020

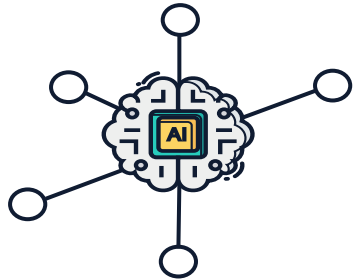


[gartner.com/SmarterWithGartner](https://gartner.com/SmarterWithGartner)

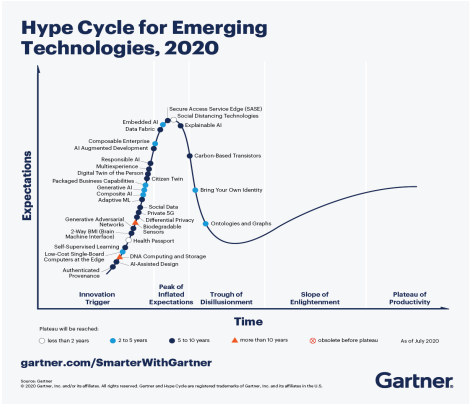
Source: Gartner  
© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.

**Gartner**

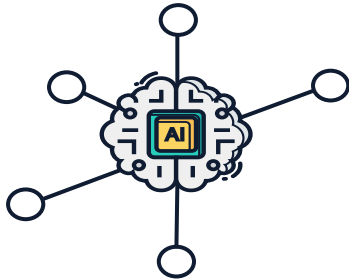
Hype vs Real?



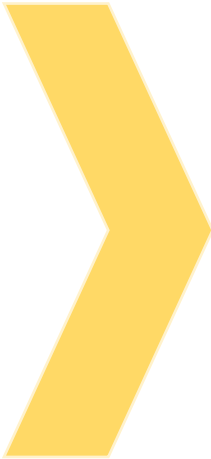
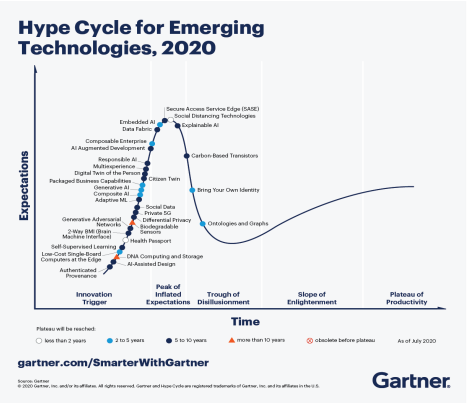
AI in finance?



Hype vs Real?



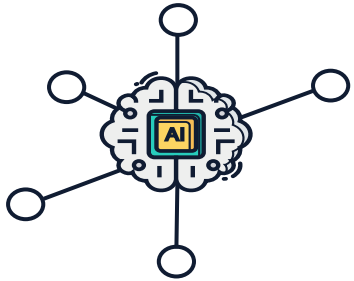
AI in finance?



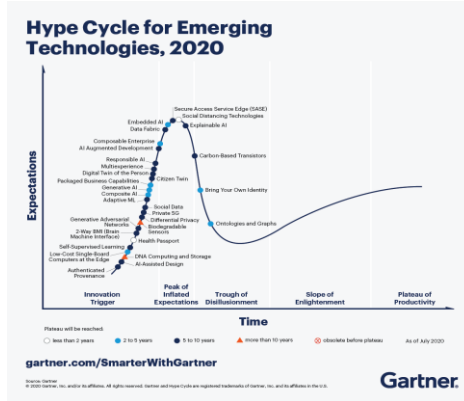
Where is the progress?



Hype vs Real?



AI in finance?



VOLUMES OF  
DATA



QUANTITATIVE  
PROBLEMS

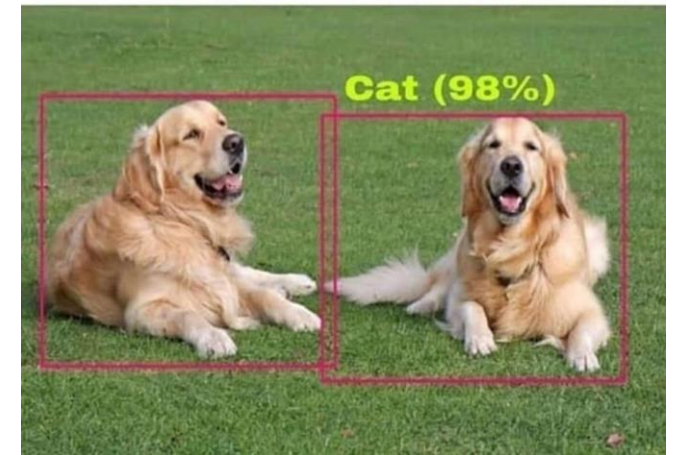
Where is the  
progress?



Well ...

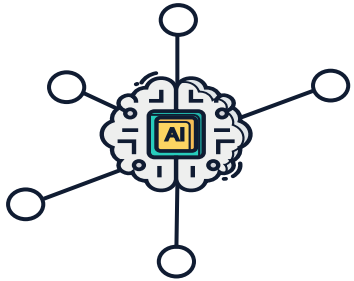
**People: \*fearing\* AI takeover**

**AI:**

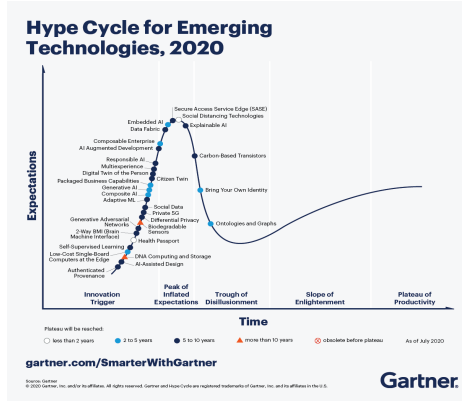


AI in practice is  
difficult

Hype vs Real?



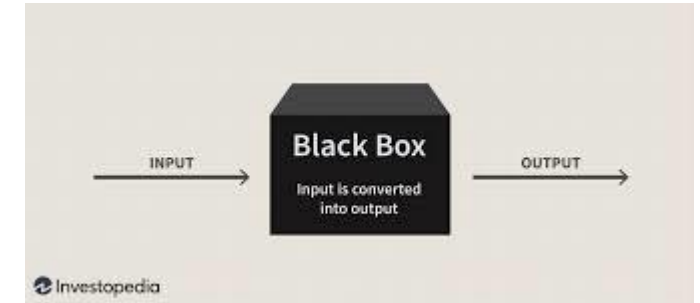
AI in finance?



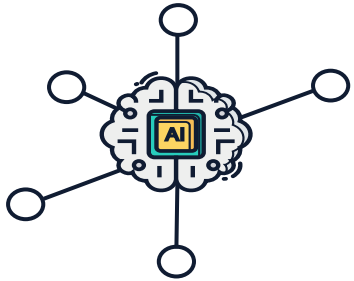
Where is the progress?



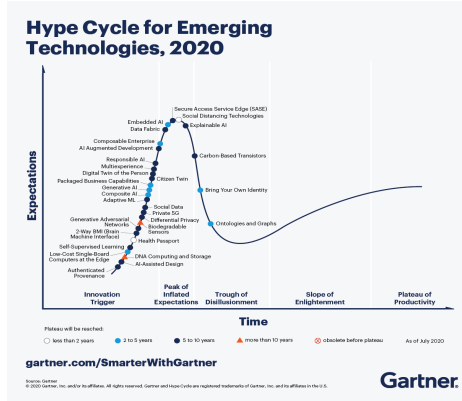
Well ...



Hype vs Real?



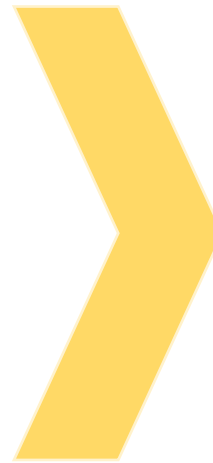
AI in finance?



VOLUMES OF  
DATA



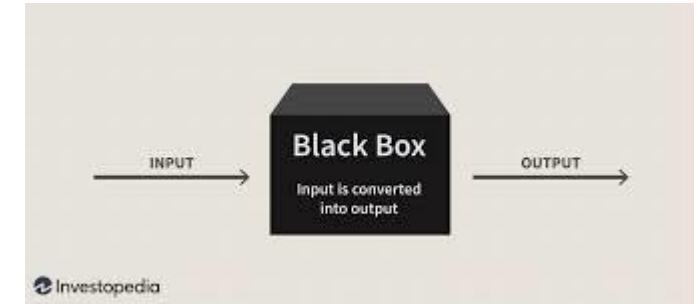
QUANTITATIVE  
PROBLEMS



Where is the  
progress?



Well ...



# Hype Cycle for Emerging Technologies, 2020



Well ...



Hype vs Real?



AI in finance?

Explainability is the name of the game!

gartner.com/SmarterWithGartner

DATA PROBLEMS

© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.

Gartner

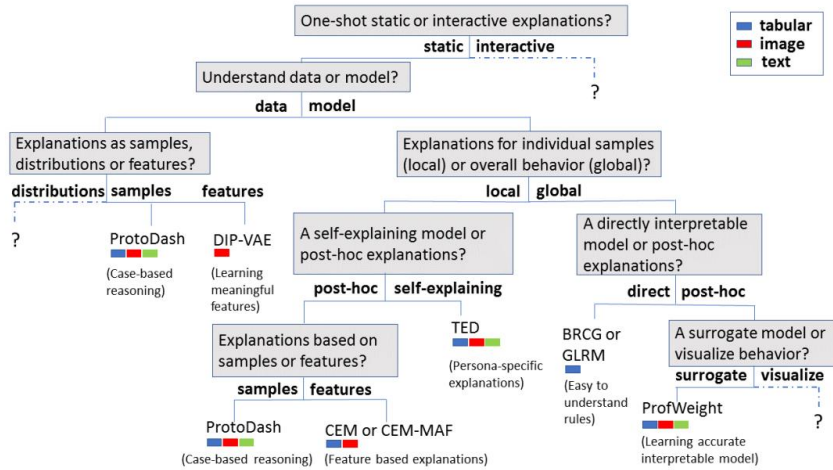


# Deploying Explainability

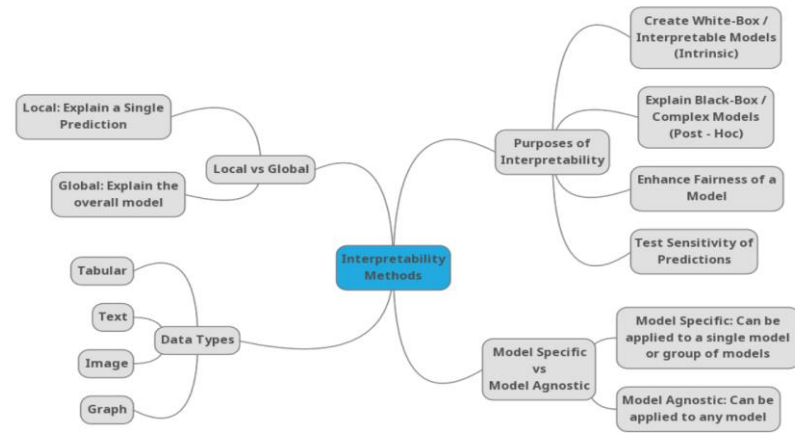
# POST-HOC Explainability

---

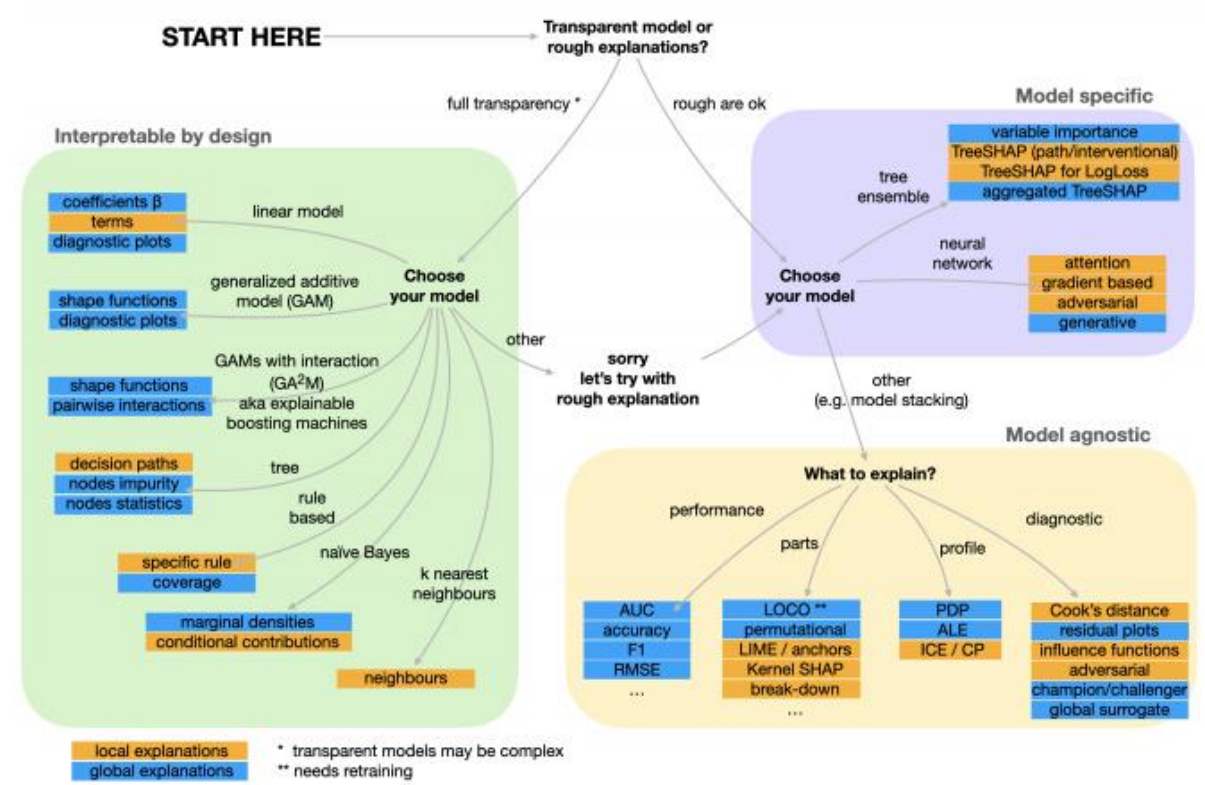
- There are many models that are highly interpretable.
- Regardless, for some ML models, **post-hoc explainability is required!**
- Post-hoc explainability techniques → understandable information about how an already developed model produces its predictions for any given input!
- Methods are considered in view of two main criteria (Linardatos et al. 2021):
  - the **type of algorithm** on which they can be applied (model-specific vs. model-agnostic)
  - **the unit being explained** (if the method provides an explanation which is instance-specific then this is a local explainability technique and if the method attempt to explain the behavior of the entire model, then this is a global explainability technique).



**Figure.** Arya et al. (2019) proposed taxonomy based on questions about what is explained, how it is explained and at what level



**Figure.** Linardatos et al. (2021) taxonomy mind-map of Machine Learning Interpretability Techniques.



**Figure.** Maksymiuk et al. (2021) model-oriented taxonomy for XAI method

# LIME: Details

---

- The explanation provided by LIME for each observation:

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where  $G$  is the class of potentially interpretable models (i.e. linear models)

$g \in G$ : An explanation considered as a model

$f: \mathbb{R}^d \rightarrow \mathbb{R}$ : The main classifier being explained

$\pi_x(z)$ : The proximity measure of an instance  $z$  from  $x$

- The goal is to **minimize the locality aware loss  $L$**  without making any assumptions about  $f$ , since a key property of LIME is that it is model agnostic.
- $L$  is the measure of how unfaithful  $g$  is in approximating  $f$  in the locality defined by  $\pi_x$ .

# Shapley Values: **DETAILS**

---

- Given a model

$$f(x_1, x_2, x_3 \dots x_n)$$

with feature 1 to  $n$  being players in a game in which the payoff  $v$  is the measure of importance of the subset.

- Marginal contribution  $\Delta_v(i, S)$  of a feature  $i$ :

$$\Delta_v(i, S) = v(S \cup i) - v(S)$$

- Let  $\Pi$  be the set of permutations of the integers up to  $N$ , and given  $\pi \in \Pi$  let  $S_{i,\pi} = \{j: \pi(j) < \pi(i)\}$  are the players preceding player  $i$  in  $\pi$ , then:

$$\phi_v(i) = \frac{1}{N!} \sum_{\pi \in \Pi} \Delta_v(i, S_{i,\pi})$$

# XAI in Credit Risk Management

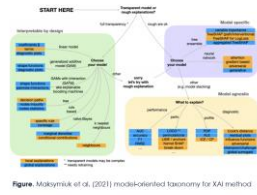
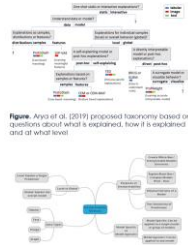
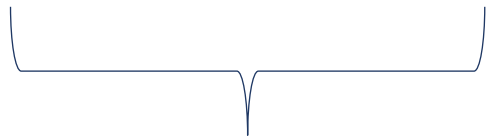
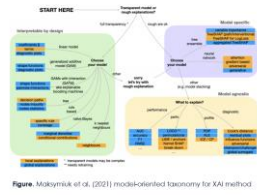
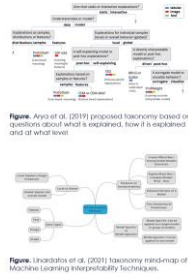


Figure 1. Ayo et al. (2019) proposed taxonomy based on questions about what is explained, how it is explained, and at what level.

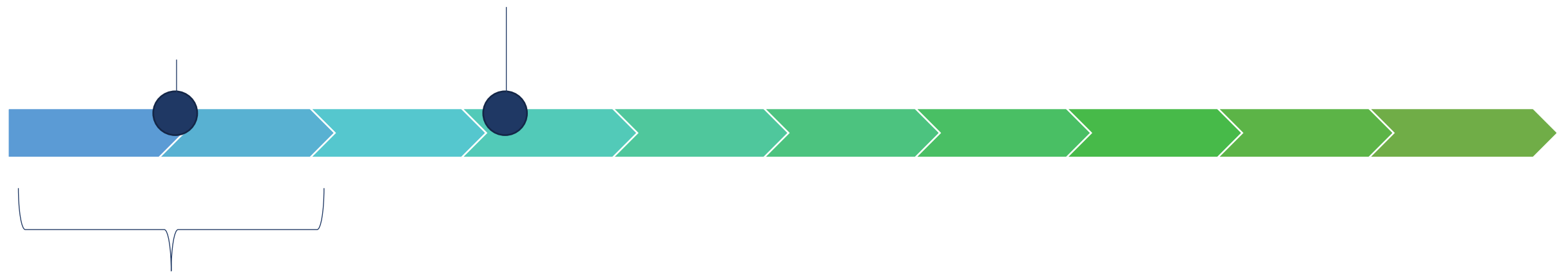
Figure 2. Mokryniuk et al. (2021) model-oriented taxonomy for XAI method.



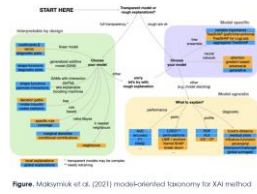
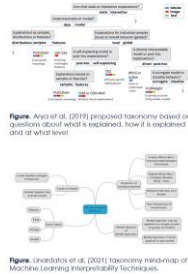
XAI Research



Match explainability  
needs of stakeholders  
with the XAI methods

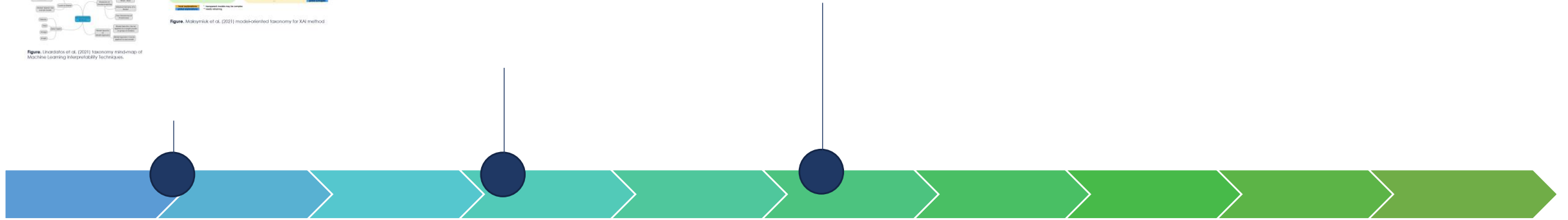


XAI Research



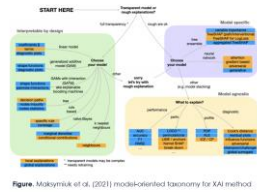
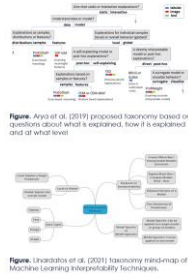
Match explainability  
needs of stakeholders  
with the XAI methods

Performance of XAI  
methods in view of  
the unique features of  
financial data



XAI Research

XAI research in **FINANCE**



Match explainability needs of stakeholders with the XAI methods

Performance of XAI methods in view of the unique features of financial data

What are the technical issues in scaling these solutions?



## Wider adoption of AI-based use cases in finance

What are the  
technical issues in  
scaling these  
solutions?

Performance of XAI  
methods in view of  
the unique features of  
financial data

Match explainability  
needs of stakeholders  
with the XAI methods

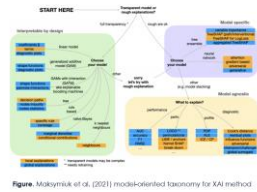
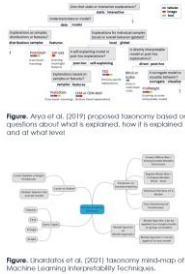


Figure 1: Ayo et al. (2019) proposed taxonomy based on questions about what is explained, how it is explained, and at what level.

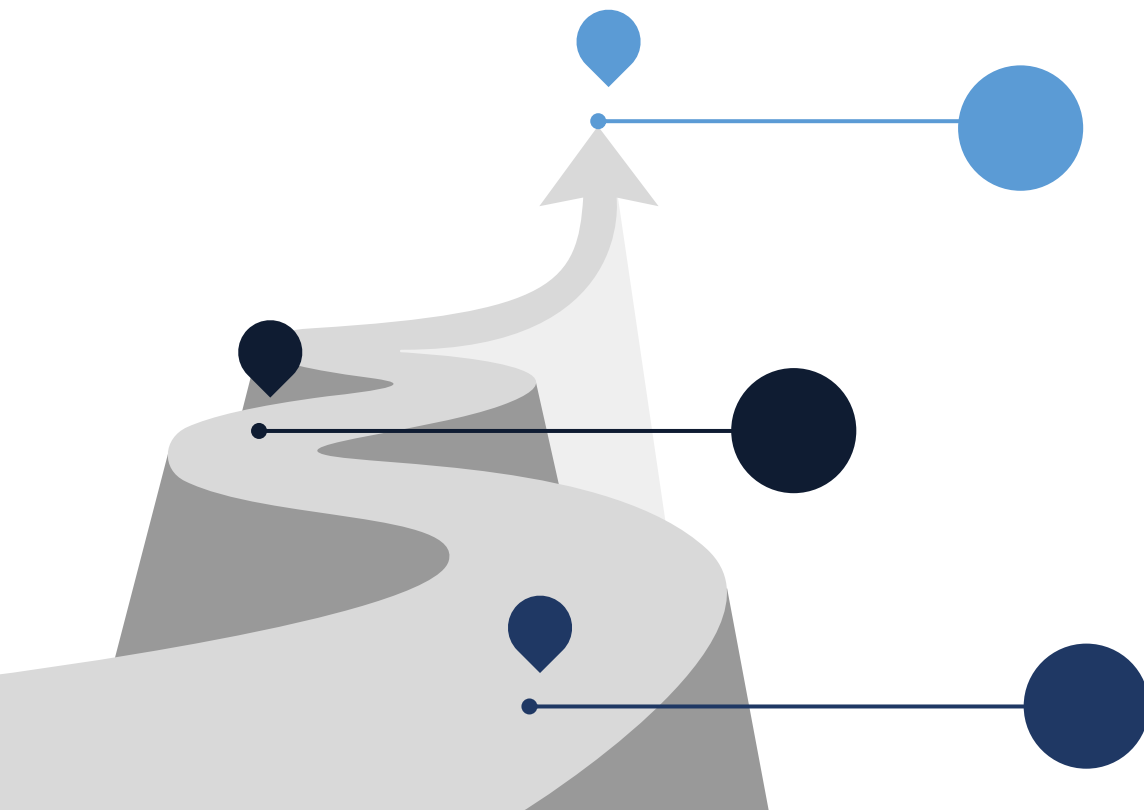
Figure 2: Makryniak et al. (2021) model-oriented taxonomy for XAI method.



# Use Case: **OBJECTIVES**

---

**Context:** Credit Risk Management



To explore the **utility of classical XAI frameworks in the context of credit risk management**

**Stability and robustness of explanations**

**Human-centric and mathematical issues**

# HUMAN-CENTRIC Issues

**Interviews**  
carried out  
with various  
stakeholders.

The main barriers for wider adoption of  
ML-based solution in finance;

The need for explainable and  
interpretable ML;

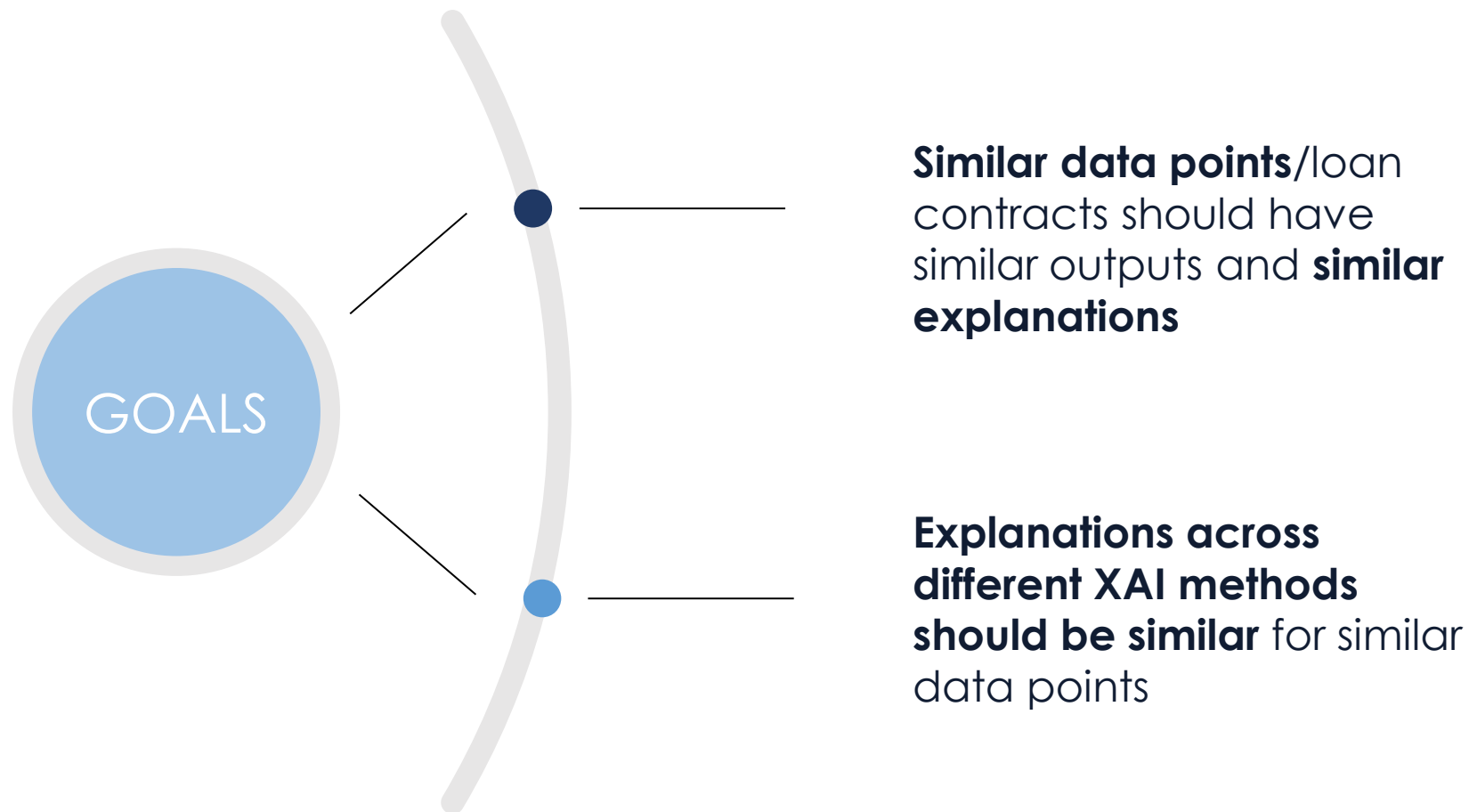
Specific explainability needs and XAI  
methods

- Explanations for **model developers**
- Could provide value for end users as well – however, **counterfactual explanations preferred**
- Visualization **not suited for end users**



# ROBUSTNESS & STABILITY of Explanations

---



# Use Case: **DATA, PROCESSING AND MODELS**



- 2GB of data and containing information [160 features] on **2.2 million loan contracts**
- Processing:
  - In order to deal with the missing values, in the first instance, all columns which had "NaN" values in more then 90% of the records, were cancelled.
  - Highly correlated features were also eliminated from the input space
  - One hot encoding and combining levels
  - Balanced target
  - Boruta algo

Table 1. Performance

Model	Parameter Space	Performance on Test Data
Logistic Regression	penalty='l2' solver='lbfgs'	Accuracy: 0.9978 , Precision: 0.9960 Recall: 0.9932, F1 score: 0.9946
XGBOOST	scoring = 'roc_auc', cv = 5, n_jobs = -1, verbose = 3, n_estimators = 100, max_depths = 4	Accuracy: 0.9971 , Precision: 1.00 Recall: 0.97, F1 score: 0.99
Random Forest	n_estimators: 500, max_depth: 20	Accuracy: 0.9932, Precision: 1.00 Recall: 0.96, F1 score: 0.98
SVM	gamma='auto', C=1.0, kernel='rbf', probability=False/True	Accuracy: 0.99487, Precision: 1.00 Recall: 0.96, F1 score: 0.98
Neural Networks	n_hidden = 2, neurons = [35,35], activations = ReLU, sigmoid loss = binary_crossentropy , Optimizer = adam	Accuracy: 0.9998, Precision: 0.9999 Recall: 0.9985, F1 score: 0.9992

# Stability of Explanations though **GRAPH THEORY**

---

- Use concepts from **graph theory** to investigate whether similar loan contracts have obtained similar explanations
- We exploit information derived from the numerical features collected in a vector  $x_n$  representing the different loan contacts  $n$ .
- We define a metric **D - standardized Euclidean distance** between each pair  $(x_j; y_j)$  loan feature vectors.

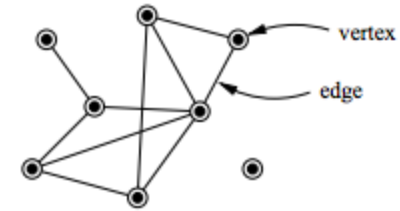


Image source: [wikipedia](https://en.wikipedia.org/wiki/File:Graph_terminology.svg)

$$D_{x,y} = \sqrt{\sum_{j=1}^J \left( \frac{x_j}{s_j} - \frac{y}{s_j} \right)^2}$$

# Stability of Explanations though **GRAPH THEORY**

---

- Use concepts from **graph theory** to investigate whether similar loan contracts have obtained similar explanations
- We exploit information derived from the numerical features collected in a vector  $x_n$  representing the different loan contacts  $n$ .
- We define a metric **D - standardized Euclidean distance** between each pair  $(x_j; y_j)$  loan feature vectors.

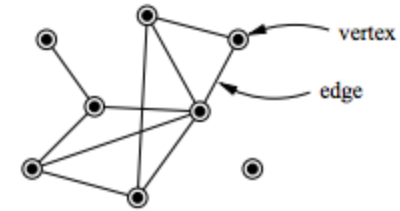


Image source: [wikipedia](https://en.wikipedia.org/wiki/File:Graph_terminology.svg)

$$D_{x,y} = \sqrt{\sum_{j=1}^J \left( \frac{x_j}{s_j} - \frac{y}{s_j} \right)^2}$$

# The Minimal Spanning Tree

---

- We derive the **Minimal Spanning Tree (MST)** representation of the loan contracts
- For a **Graph  $G$** , the goal is to find a tree  $T$  which is a spanning subgraph of  $G$ , i.e. every node is included to at least one edge of  $T$  and has minimum total weight.
  - Pick some arbitrary start node  $u$ . Initialize  $T = u$
  - At each step add the lowest-weight edge to  $T$  (the lowest-weight edge that has exactly one node in  $T$  and one node not in  $T$ );
  - Stop when  $T$  spans all the nodes.

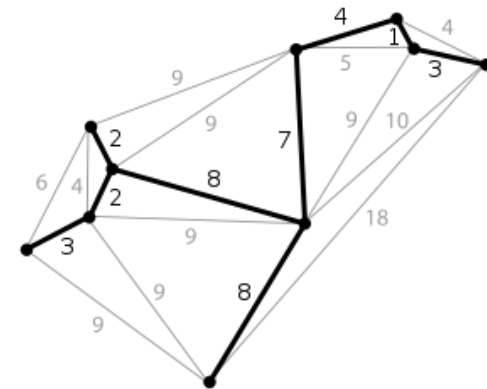
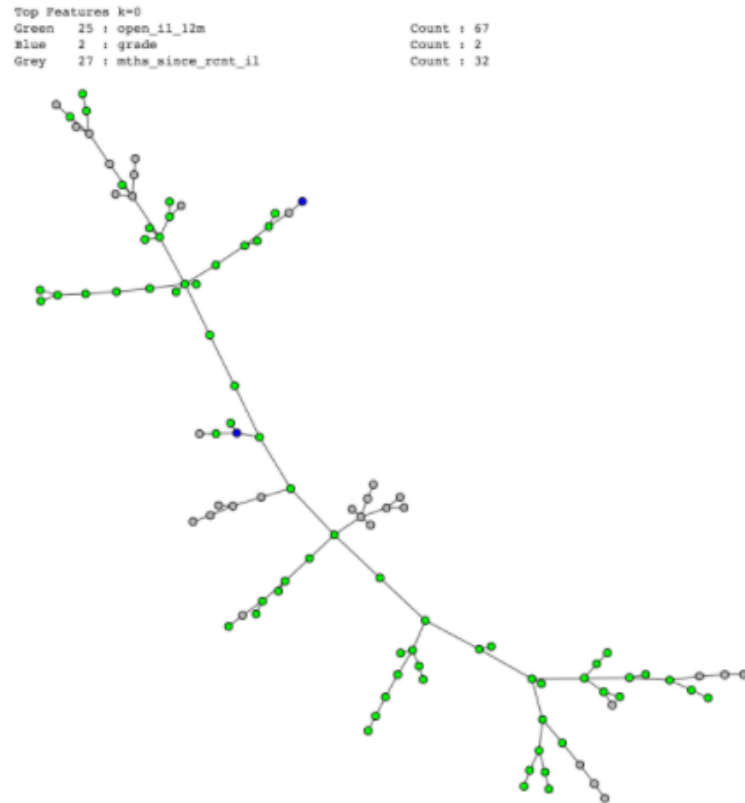
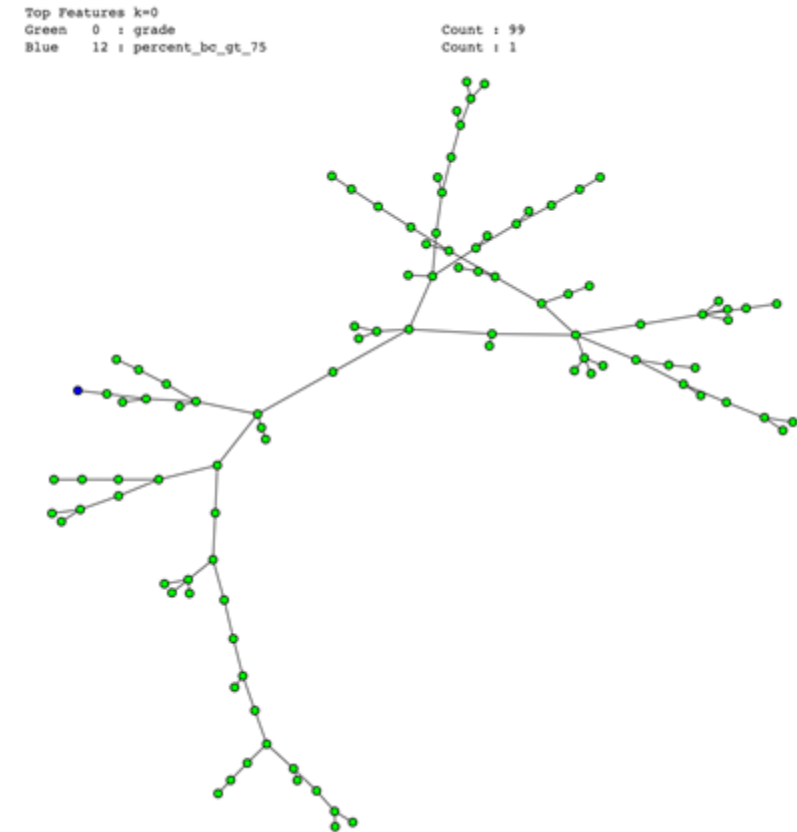


Image source: [wikipedia](https://en.wikipedia.org/wiki/Minimal_spanning_tree)

# Stability of Explanations though **GRAPH THEORY**

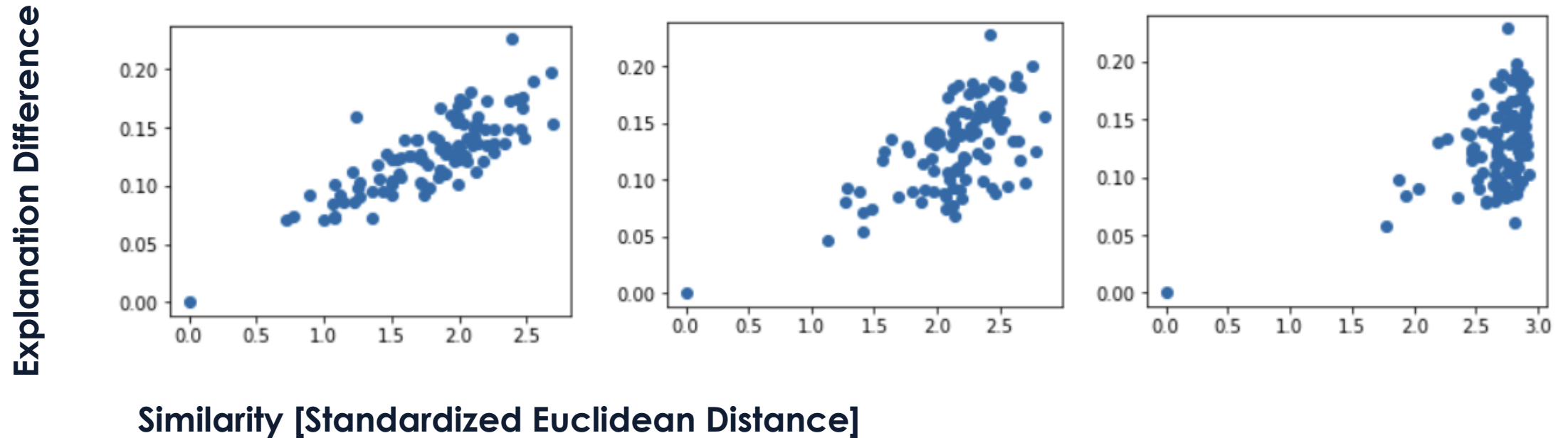


**Figure.** MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = “Number of instalment accounts opened in past 12 months”; grey = “Months since most recent instalment accounts opened” ; blue = “Grade”]



**Figure.** MST tree representation of 100 random data points. Coloring based on the top explanatory feature [green = “Grade”, blue = “Percent of trades never delinquent”]

# Stability of Explanations though **GRAPH THEORY**



**Figure.** Explanation Difference vs Spatial Distance for  $\text{ref}_i = 1000$ ,  $n = 100$  for 5, 10, and 20 Features.

\*The Explanation Difference formula takes the top  $n$  features of two points, adds up the squared difference of the contributions of each feature in common, and for each feature that is not common, adds up the square of each contribution then finally take the square root of the sum.

# TECHNICAL Issues

---

- Issue with the **different estimation procedures**
  - the exact computation of the Shapley value is computationally intensive
  - Feature selection can be crucial
    - The choice of features that count as players can affect the resulting explanations
- Only few model-specific solutions for the computational complexity
- Classical approaches and their current implementation are not tailored for **financial time series** (subject to trends, vola-clusters,...).
- Specifically, **perturbation-based methods are fully dependent on the ability to perturb samples in a meaningful way**. In the context of financial data:
  - if features are correlated, the artificial coalitions created will lie outside of the multivariate joint distribution of the data,
  - if the data are independent, coalitions can still be meaningless;
  - generating artificial data points through random replacement disregards the time sequence hence producing unrealistic values for the feature of interest.

# Outlook & Follow-up Work - I

---

- The **lack of algorithmic transparency is one of the main barriers** for the wider adoption of AI-based solutions in credit risk management
- **Two-fold objective** of the work:
  - human-centric and mathematical issues related with the implementation of XAI methods in finance, and
  - explore the stability and robustness of explanations provided
- **Human-centric issues** → we find that that XAI methods are suited to the needs of ML engineers
- **Stability** → state-of-art methods offer certain level of stability
- **Technical issues** → many challenges

# Outlook & Follow-up Work - II

- Need for a **time series approach to explainability**
- Emphasize input-output relations instead → **infinitesimal changes**
- Partial derivative of:
  - Net-output with respect to explanatory (input) variables:  $\frac{\partial o_t}{\partial x_{it}}$
  - X-function  $xf(o_t)$  of the net-output with respect to the input:  $\frac{\partial xf(o_t)}{\partial x_{it}}$
- Interpretability/XAI: **observe sensitivities** (gradient/derivative of X-function) **over time**

