Universidad de Navarra

DATAI
INSTITUTO DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

II JORNADAS
de Ciencia de los Datos e Inteligencia Artificial

9 AL 12 DE MAYO DE 2022

Aula Siemens Gamesa
Edificio Ismael Sánchez Bella
Universidad de Navarra
Pamplona

INFORMACIÓN GENERAL:

El Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI) es un centro de investigación, innovación y formación de la Universidad de Navarra que busca un trabajo colaborativo de diferentes grupos y personas de la Universidad con el objetivo de alcanzar la excelencia en Ciencia de Datos e Inteligencia Artificial.

En sus objetivos existe un importante apoyo en la investigación, así como en innovación y transferencia de conocimientos al ámbito industrial, empresarial y social. También existe una fuerte implicación en la formación de investigadores, estudiantes y profesionales. Por esta razón, el Instituto dirige un Máster oficial en Big Data Science destinado principalmente a jóvenes profesionales, así como estudiantes de doctorado, en diferentes áreas con necesidades importantes de análisis de datos.

En estas segundas jornadas 25 investigadores expondrán trabajos de investigación en los que están trabajando en estos momentos. Todos ellos muestran la transversalidad del Instituto.

II JORNADAS DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

del 9 al 12 de mayo 2022

Web

Dirección:

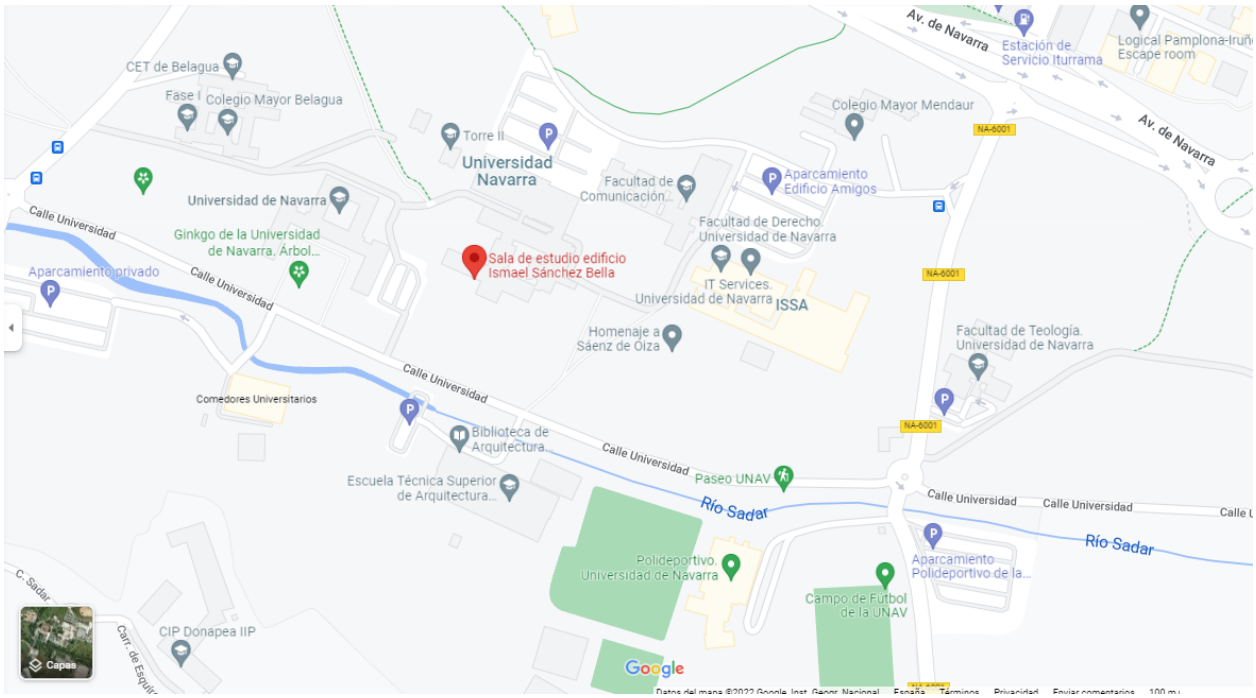
AULA SIEMENS GAMESA - Planta Baja.

Edificio Ismael Sánchez Bella

UNIVERSIDAD DE NAVARRA

PAMPLONA

Mapa:



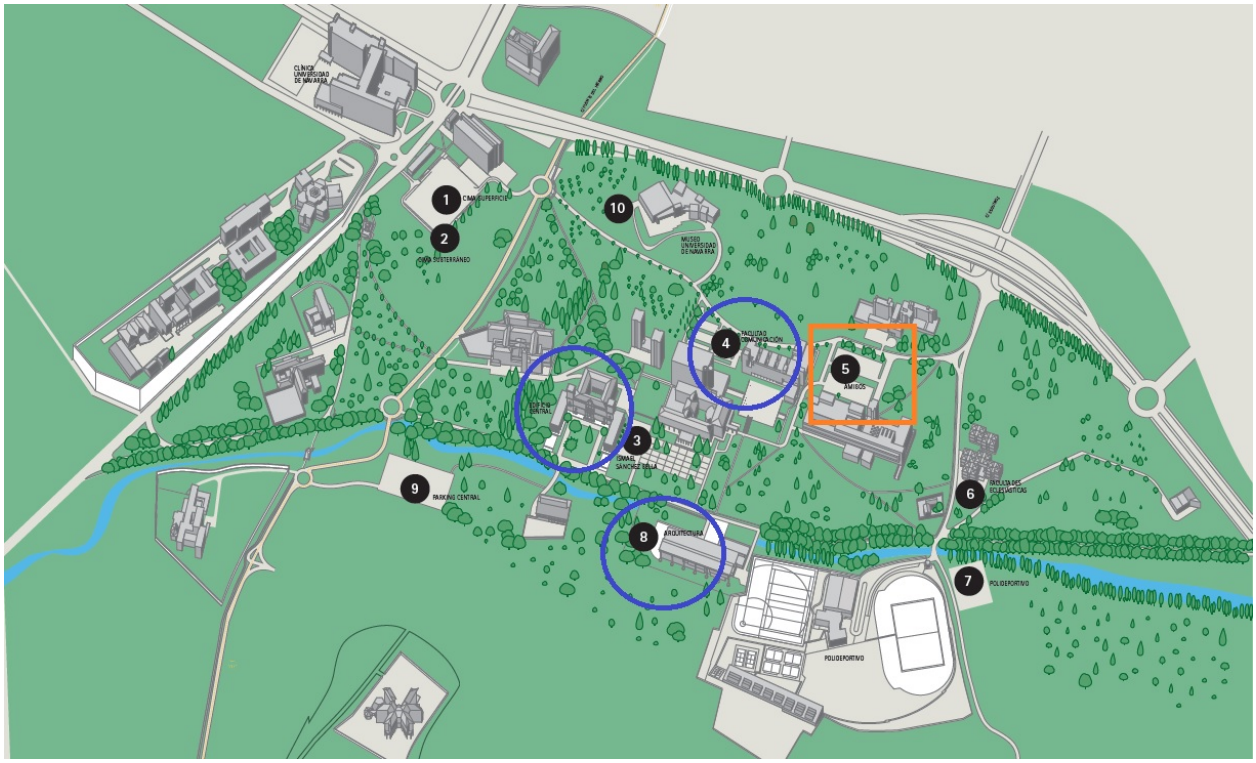
ALOJAMIENTO:

Hotel Blanca de Navarra****	**NH Pamplona Iruña Park ******	**Hotel Sancho Ramirez *****
WEB Tel. +34 948 171 010 Avd. Pío XII, 43. 31008 Pamplona (Navarra) reservas@hotelblancadenavarra.com	WEB nhirunapark@nh-hotels.com C/ de Arcadio María Larraona, 1, 31008 Pamplona, Navarra Tel. +34 948 197 119	WEB C/ Sancho Ramírez, 11. 31008 Pamplona Tel. +34 948. 271 712

COMIDAS:

Restaurantes dentro de la Universidad:

- Edificio Central
- Edificio de Comunicación
- Edificio Arquitectura



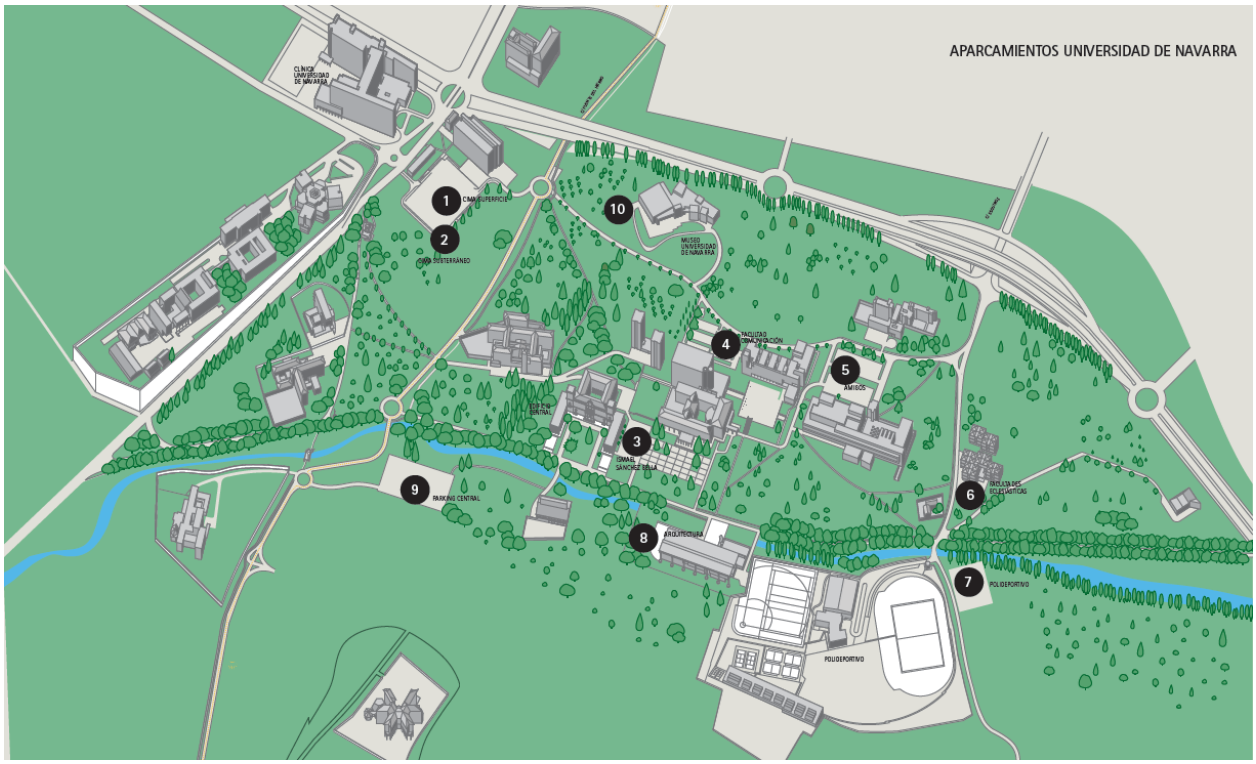
Restaurantes cercanos fuera de la Universidad:

Casa Amparo WEB Calle Esquiroz, 22 bajo Trasera. 31007 Pamplona, Navarra. Teléfono: 948 261 162	**Restaurante Aragón** WEB C. Buenaventura Íñiguez, 2, 31006 Pamplona, Navarra Teléfono: 948 23 30 96
Restaurante Aragón WEB C. Buenaventura Íñiguez, 2, 31006 Pamplona, Navarra Teléfono: 948 23 30 96	**Casa Luis** WEB Dirección: C. Padre Calatayud, 11, 31003 Pamplona, Navarra Teléfono: 948 15 18 21

APARCAMIENTO EN EL CAMPUS:

Si desea reservar aparcamiento en la universidad, cumplimente el siguiente formulario indicando el día o los días que desea reservar.

Los coches no pueden pernoctar en la universidad. Hay que sacarlos del parking por la noche.



Comité Científico:

- Amparo Alonso Betanzos
- Enrique del Castillo
- Nuria Oliver
- John Stufken
- Trevor Hastie (Honorary Member)

Comité Organizador:

- Jesús López Fidalgo
- Stella Salvatierra
- Elisabeth Viles Diez
- Sergio Ardanza-Trevijano
- José Miguel Leiva Murillo

PROGRAMA

LUNES, 9 DE MAYO		
11.00 - 12.00	**REGISTRO**	
12.00 - 12.45	**APERTURA DE LAS JORNADAS**	• **María Iraburu.** ***Rectora. Universidad de Navarra*** • **Pilar Concejo.** ***Directora de Formación y Gestión del Talento en el Grupo BBVA*** • ***Jesús López Fidalgo. Director de DATAI***
12.50 - 13.50	**PLENARIA a cargo de Pedro Larrañaga** **Modera: Rubén Armañanzas**	***Redes bayesianas en modelización y optimización. De la neurociencia a la industria 4.0***

MARTES, 10 DE MAYO		
SESIÓN 1 Modera: Jean Bragard		
9.20 - 9.45	**Leire Moriones**	***Using High-resolution voltage maps to predict "redo" in the treatment of atrial fibrillation (AF)***
9.45 - 10.10	**Sergey Yablonsky**	***AI-DRIVEN ORGANIZATION MATURITY: A CONCEPTUAL FRAMEWORK***
10.10 - 10.35	**Javier Porras Castaño**	***Dicho y hecho: la IA ya permite democratizar la programación a cualquier perfil***
10.35 - 11.00	**Iñaki Fernández de Troconiz**	***Characterizing the interplay between biological factors and cell growth in unperturbed tumor growth dynamics***
11.00 - 11.30	**CAFÉ**	
11.30 - 12.20	**PLENARIA: José Antonio Lozano** **Modera: Sergio Ardanza**	***Test de hipótesis: sus puntos débiles y alternativas Bayesianas.***
SESIÓN 2 Modera: Ignacio Rodríguez		
12.30 - 12.55	**Paloma Marín Martínez**	***Model-driven or data-driven? Pricing and prediction in multi-dealer to client platform for european government bonds***
12.55 - 13.20	**Christian Ojeda Trejo**	***Optimal market making under information asymmetry***
13.20 - 13.45	**Conrado Javier García Montiel**	***Multi-product dynamic pricing strategies with unlimited inventories***

MIÉRCOLES, 11 DE MAYO		
SESIÓN 3 Modera: Mikel Hernáez		
9.20 - 9.45	**Luis Manuel García Muñoz**	***Machine Learning Techniques for Financial Market Risk measurement***
9.45 - 10.10	**Adrián Díaz Lanchas**	***Fraud detection based-on Web browsing behavioral patterns***
10.10 - 10.35	**Priscila Espinosa Adamez**	***Automatic tools for measuring the economic regional growth***
10.35 - 11.00	**Juan José Fernández tebar**	***Bundle Pricing of Cash Management Products with Mixed Integer Programming***
11.00 - 11.30	**CAFÉ**	
11.30 - 12.20	**PLENARIA: Antonio García Marqués** **Modera: José Miguel Leiva**	***Connecting the dots: learning graphs from nodal signal observations***
SESIÓN 4 Modera: Jorge Elorza		
12.30 - 12.55	**Sergio Ardanza-Trevijano**	***Detecting cancer-specific copy number changes using topological data analysis***
12.55 - 13.20	**Idoia Cerro Belzuz**	***Sistema sensorial con inteligencia artificial para guantes sensoriales***
13.20 - 13.45	**Francisco Javier Talavera Andújar**	***On the computation of triangular norms in chains***

JUEVES, 12 DE MAYO		
SESIÓN 5 Modera: María Fernández Seara		
9.20 - 9.45	**Matias Ávila Clemente**	***Modeling multiple seasonalities of NO2 hourly pollution levels***
9.45 - 10.10	**Miguel A. Alcantara Duran**	***Evaluation of pollutants and particulate material in interior spaces***
10.10 - 10.35	**Yury Jiménez Agudelo**	***Mathematical model and characterization of emotional states in bipolar disorders***
10.35 - 11.00	**Mario Martínez García**	***Learning a battery of COVID-19 mortality prediction models by multi-objective optimization***
11.00 - 11.30	**CAFÉ**	
11.30 - 12.20	**PLENARIA: Daniel Peña** **Modera: Stella Salvatierra**	***Some Recent Methods for Analyzing High Dimensional Time Series***
SESIÓN 6 Modera: Iñaki Fernández de Trocóniz		
12.30 - 12.55	**Jose Antonio Moler Cuiral**	***Designing experiments for estimating an appropriate outlet size for a silo type problem***
12.55 - 13.20	**Álvaro Cía Mina**	***Optimal subdata selection assuming random covariates***
13.20	**CLAUSURA**	

ABSTRACTS

CONFERENCIAS PLENARIAS:

Pedro Larrañaga Múgica
Universidad Politécnica de Madrid

REDES BAYESIANAS EN MODELIZACIÓN Y OPTIMIZACIÓN. DE LA NEUROCIENCIA A LA INDUSTRIA 4.O

La charla proporcionará una introducción intuitiva a las redes Bayesianas como modelos interpretables del aprendizaje automático con las que llevar a cabo distintos tipos de razonamiento probabilístico tanto predictivo como diagnóstico, intercausal o contrafactual. Veremos ejemplos de aplicación de dichos modelos en diversos problemas de modelización en dominios como la neurociencia computacional o la industria 4.0, tanto en problemas estáticos como continuos. En optimización se introducirán los algoritmos de estimación de distribuciones, una técnica heurística basada en la evolución de poblaciones por medio del aprendizaje y la posterior simulación de redes Bayesianas que modelan el comportamiento de los mejores individuos de cada generación.

Jose Antonio Lozano
Basque Center For Applied Mathematics BCAM

TEST DE HIPÓTESIS: SUS PUNTOS DÉBILES Y ALTERNATIVAS BAYESIANAS

El uso de test de hipótesis en el ámbito científico ha sufrido numerosos vaivenes en las últimas dos décadas: desde un auge importante donde cada trabajo experimental publicado debía llevar asociado un estudio estadístico basado en los mismos, hasta un rechazo absoluto donde varias revistas científicas prohibieron la publicación de trabajos que contuviesen test de hipótesis como método de validación de los resultados experimentales. En esta charla analizaremos inicialmente los puntos fuertes y débiles de los test de hipótesis, exponiendo el porqué de dicha controversia. Tras ello, en una segunda parte, propondremos alternativas Bayesianas para realizar análisis de datos experimentales más elaborados y ricos que los proporcionados por los test de hipótesis.

Antonio García Marqués
Universidad Rey Juan Carlos

CONNECTING THE DOTS: LEARNING GRAPHS FROM NODAL SIGNAL OBSERVATIONS

The talk will provide an overview of graph signal processing (GSP)-based methods designed to learn an unknown network from nodal observations. Using signals to learn a graph is a central problem in network science and statistics, with results going back more than 50 years. The main goal of the talk is threefold: i) explaining in detail fundamental GSP-based methods and comparing those with classical methods in statistics, ii) putting forth a number of GSP-based formulations and algorithms able to address scenarios with a range of different operating conditions, and iii) briefly introducing generalizations to more challenging setups, including multi-layer graphs and learning in the presence of hidden nodal variables. Our graph learning algorithms will be designed as solutions to judiciously formulated constrained-optimization sparse-recovery problems. Critical to this approach is the codification of GSP concepts such as signal smoothness and graph stationarity into tractable constraints. Last but not least, while the focus will be on the so-called network association problem (a setup where observations from all nodes are available), the problem of network tomography (where some nodes remain unobserved, and which can be related to latent-variable graphical lasso) will also be discussed.

Daniel Peña Sánchez de Rivera
Universidad Carlos III de Madrid

SOME RECENT METHODS FOR ANALYZING HIGH DIMENSIONAL TIME SERIES

This presentation will describe three recent advances that are useful for the analyses of high dimensional time series. The first one concerns new ways for visualizing large sets of time series. Dynamic quantiles will be introduced and some multivariate plots will be shown illustrating ways to reveal the information in the set of time series. The second one is clustering time series by dependency. A new way to cluster the series by their linear dependency will be described and shown to be able to split the set of series into homogeneous groups. The third one is dynamic factor models, and some examples of

high dimensional tensor factor models will be given. The future evolution of the field of dependent high dimensional data will be discussed in the conclusions.

COMUNICACIONES:

EVALUATION OF POLLUTANTS AND PARTICULATE MATERIAL IN INTERIOR SPACES

Alcántara MA[1], Santamaría C.[1] , Martín-Gómez, C.[2]

[1] Department of Chemistry, University of Navarra

[2] Department of Construction, Installations and Structures, University of Navarra

There exist gaps with respect to the long-term problems generated by pollutants within non-industrial interior spaces. Human beings spend approximately 80% of their time in these spaces, a fact that has been aggravated due to increasingly hermetic constructions, which fundamentally aim to reduce energy expenditure, limiting the exchange rate of external air. Other factors that also influence indoor air quality are construction materials, decoration, cleaning products, and external and internal activities. As a result, a cocktail of polluting substances is produced that makes it difficult to know their composition, interaction, and daily dose exposure (Guardino, 2010).

In a non-industrial indoor space, between 50 and 300 different volatile organic compounds (VOCs) can be present, as well as SO_2, NOx, NO_2, particulate matter (PM), among others (Ruiz & García Sanz, 2010). The presence of these compounds can cause a decrease in work performance, and some studies show that a large part of these are irritants of the mucous membrane, eyes, skin and part of them are suspected or proven CMR (carcinogenic, mutagenic and/or toxic to reproduction (Marta Morales, Blanco Acevedo, & García Nieto, 2010)

The objective of this project is to determine the exposure and levels of pollutants in our work, study and home environment. Therefore, the research will be carried out within the university campus, considered as a small city where (8) buildings with different characteristics and multiple activities will be analyzed, which can be extrapolated to any city environment.

USING TOPOLOGICAL SIGNATURES TO DETECT CANCER SPECIFIC COPY NUMBER CHANGES

Sergio Ardanza-Trevijano, Jai Aslam, Jingwei Xiong, Javier Arsuaga and Radmila Sazdanovic

Universidad de Navarra

Copy number changes play an important role in the development of cancer and are commonly associated to changes in gene expression. Persistence curves, such as Betti curves, have been used to detect copy number changes, however it is known these curves are unstable with respect to small perturbations in the data. We address the stability of lifespan and Betti curves by providing bounds on the distance between persistence curves of Vietoris-Rips filtrations built on data and slightly perturbed data in terms of the bottleneck distance. Next, we perform simulations to compare the predictive ability of Betti curves, lifespan curves (conditionally stable) and stable persistent landscapes to detect copy number aberrations.

We use these methods to identify significant chromosome regions associated with the four major molecular subtypes of breast cancer: Luminal A, Luminal B, Basal and HER2 positive. Identified segments are then used as predictor variables to build machine learning models which classify patients as one of the four subtypes. We find that no single persistence curve outperforms the others and instead suggest a complementary approach using a suite of persistence curves.

MODELLING MULTIPLE SEASONALITIES OF NO2 HOURLY POLLUTION LEVELS

M. Avila[1], A.M. Alonso[2], D. Peña[3]

[1] Department of Statistics, University Carlos III of Madrid
[2] Department of Statistics and Institute Flores de Lemus
[3] Department of Statistics and UC3M-Santander Big Data Institute

NO2 is one of the most common pollutants in urban areas and road traffic is the main source of this contaminant. The NO2 hourly time series has three seasonal patterns due to human activity and climatological conditions. The classical approach assumes that those seasonalities are determinist and can be modelled by trigonometric functions or dummy variables. This assumption may be too strict. A more flexible model is to allow the seasonality to slowly change as in a seasonal ARIMA model, where the seasonality is modelled as a stochastic processes. In this paper, we propose to model them iteratively combining different seasonal ARIMA models.

We proposed a model that takes into account the regular dependency between hourly observations, the three seasonal components (daily, weekly and annual with seasonality 24, 168 and 8736 hours, respectively) and covariables such as the current average wind speed. It is worth noting that this model is not linear since the seasonal component is composed by varying parameters depending on the day hour and weekday. In order to estimate the model we have performed an approximation in two sequential steps: (1) In the first step we stratify the hourly NO2 time series into 168 weekly time series, formed by each of the hours of the day and the days of the week. Each of the 168 weekly subseries is modelled separately with a seasonal ARIMAX52 and covariables. The regular component of this ARIMAX52 model captures the weekly seasonality while the seasonal one captures the annual seasonality. (2) Secondly, we consider the residuals from the first step in their natural order and fit a seasonal ARIMA24. The seasonal component of this ARIMA24 model will capture de daily seasonality while the regular component will capture the dependency between an observation and the immediately preceding ones. We compare our approach with other methods that have been developed to consider more seasonalities such as TBATS and Prophet, where the seasonal components are modelled by trigonometric functions.

SMART SYSTEM WITH ARTIFICIAL INTELLIGENCE FOR SENSORY GLOVES

Idoia Cerro[1] , Iban Latasa[1,2] , Claudio Guerra[3] , Pedro Pagola[2,4]
Blanca Bujanda[2,4] and José Javier Astrain[2,5]

[1] IED Electronics, Pol. Ind. Plazaola, E6, 31195 Berrioplano, Spain
[2] Department of Statistics, Computer Science and Mathematics, Public University of Navarre,
[3] Plant Pamplona SAS Autosystemtechnik, S.A., Faurecia, Polígono Industrial de Arazuri-Orcoyen,
[4] INAMAT2-Institute for Advanced Materials and Mathematics, Public University of Navarre,
[5] Institute of Smart Cities, Public University of Navarre, 31006 Pamplona, Spain

This paper presents a new sensory system based on advanced algorithms and machine learning techniques that provides sensory gloves with the ability to ensure real-time connection of all connectors in the cabling of a cockpit module. Besides a microphone, the sensory glove also includes a gyroscope and three accelerometers that provide valuable information to allow the selection of the appropriate signal time windows recorded by the microphone of the glove. These signal time windows are subsequently analyzed by a convolutional neural network, which indicates whether the connection of the components has been made correctly or not. The development of the system, its implementation in a production industry environment and the results obtained are analyzed.

OPTIMAL SUBSAMPLING ASSUMING RANDOM COVARIATES

Cía Mina, Álvaro[1,2], López Fidalgo, Jesús[1,2]

[1] Universidad de Navarra, Institute of Data Science and Artificial Intelligence (DATAI)

[2] Universidad de Navarra, Tecnun Escuela de Ingeniería

The subsampling procedure is widely used to downsize the data volume and allows computing estimators in regression models. Usually, subsampling is performed defining a weight for each point and selecting a subset according to these weights. The subsample can be chosen at random (Passive Learning), but in order to obtain better estimators, the optimal experimental design theory can be used searching for an influential sub-sample (Active Learning). This has been developed in the literature for linear and logistic regression, obtaining algorithms based on D-optimality and A-optimality. To the authors knowledge the distribution of the explanatory variables has never been considered for obtaining a subsample. We study the effect of the explanatory variables distribution on the estimation as well as the optimal design. We propose a novel method to obtain optimal subsampling, taking into account the marginal distribution of the covariates.

FRAUD DETECTION BASED-ON WEB BROWSING BEHAVIOURAL PATTERNS

Díaz Lanchas, Adrián

BBVA

In recent years, the unprecedented growth in digital services in the banking sector has brought about changes in fraud and illicit activity. These changes have caused most classical Fraud Prevention Systems to be less capable. One of the reasons for this decrease in fraud detection might be that most of the systems are based mainly on the analysis of transactional data of the client and the operation itself, without considering the interaction and behaviour of the user with the digital services. In this work, we propose to detect illicit transfers by analysing the behavioural web browsing patterns of the client in the BBVA website. For this purpose, instead of using navigation aggregated data, we characterize the activity of the client and all the transitions (steps from one page to another) to build a navigation graph (sequence that contains all of the pages and actions of the user inside the customer portal). With this graph, each navigation is described as an adjacency matrix through which we apply supervised classification based on neural networks. When developing this solution we have faced three main issues: extreme class imbalance, high-dimensionality and demanding business requirements. At first, data classes were extremely imbalanced (only 0,0002% fraudulent sessions) and, consequently, it was necessary to rely on sampling techniques to avoid underfitting. Second, the high number of pages caused a large number of features so a reduction had to be applied. Finally, from a business perspective, both types of classification errors were costly: undetected fraudulent activity meant customers money was lost while missdetections generated a poor customer experience. First results showed that, over a test sample and using imbalanced classification in the training phase, more than 50% of fraud was detected and 99.9% of legitimate navigation was correctly classified. If we also applied random balanced classification in the training phase, more than 92% of fraudulent transactions were detected in the test.

AUTOMATIC TOOLS FOR MEASURING THE ECONOMIC REGIONAL GROWTH

Espinosa Adamez, Priscila
Universidad de Valencia

Tools for automating economic forecasting. A special emphasis on the monthly synthetic economic indicator.

Currently, one of the main problems economic agents must face is decision-making in environments of uncertainty. Since the last economic crisis, the current pandemic, the number and origin of economic-financial uncertainties have increased in magnitude and intensity, as is happening with the Ukrainian invasion. Regional economic agents are avid of mechanisms capable of synthetically showing the economic situation that the regional economy is going through. This paper presents the experience of the Valencian Community (Spain) in the generation of automatic tools with Shiny for measuring regional economic growth in the short, medium and long term, to improve decision-making by agents in the face of possible changes in the economic scenario.

Keywords: Economic forecasting, shiny, times series, dynamic models and R.

BUNDLE PRICING OF CASH MANAGEMENT PRODUCTS WITH MIXED INTEGER PROGRAMMING

Fernández Tebar, Juan José

BBVA

Pricing is a crucial aspect of the banking sector and is closely related to many corporate financial products.Data and advanced analytics can help optimize pricing by injecting science into decisions. The banking sector has grown into massively complex technology firms that operate a highly sophisticated network of financial markets, credit markets, and payment systems. For this reason, simplifying the buying experience is fundamental. The purpose of this research is to explore a bundle approach to pricing financial products in a competitive environment, specifically customized bundles, where a financial company defines several bundle prices that depend on the number of products and units sold for each financial product. That makes the experience of purchasing your product efficient and straightforward. This work formulates a nonlinear mixed-integer model considering business constraints, targeted at maximizing the company's revenue, automating processes, saving costs, and reducing the level of service to unprofitable clients.

Keywords: Pricing, Product Bundle Pricing, Pricing Strategy, Mixed-Integer Programming, Process Automation.

MULTI-PRODUCT DYNAMIC PRICING STRATEGIES WITH UNLIMITED INVENTORIES

García Montiel, Conrado J.

BBVA

Dynamic pricing is a pricing strategy in which prices for products and/or services are set based on market demands. Research in this area has mainly focused focuses on the single-product case with limited inventories. In this Doctoral Thesis, I would like to analyse a multi-product dynamic pricing problem with unlimited inventories. Multi-product refers to the customer portfolio in GTB (Global Transaction Banking) including Cash \& Liquidity Management, Working Capital, Securities Services and Trade Finance products. This optimization problem relaxing the inventory restriction will result in customized demand curves.

Thus, by moving pricing strategy from a mechanical reaction to a strategic instrument, BBVA will have the opportunity to capture greater value in a changing market. These strategies will be fed with internal data (historical price \& quantity), external data (financial statements, micro \& macro variables) and alternative data (regulatory requirements, sentiment, covid period...).

In this context, a robust customer segmentation based on its value will be needed. Customer value can be classified into groups based on its churn probability, current value and potential value. Acquiring and retaining the most profitable customers is a key challenge that needs to be addressed.

The outcome will be the portfolio’s optimized weights (array of prices and quantity) which maximizes the revenue. In addition to BBVA’s revenue we will be creating value for our customers through this customer-centric pricing. Forgetting about numerous negotiations of single products without taking into account the customer context. It will become a decisive factor in banking profitability.

MACHINE LEARNING TECHNIQUES FOR FINANCIAL MARKET RISK MEASUREMENT

García Muñoz, Luis Manuel

BBVA

Market risk is the risk of losses in financial portfolios due to changes in market variables (equity shares, interest rates, volatilities, ...). Financial portfolios usually contain financial derivatives such as futures and options. The valuation of some of these derivatives is computationally expensive. Since market risk measurement involves calculating the impact on the value of these portfolios under a high number of scenarios, the computational burden can be prohibitive in some cases. In this seminar we will see how supervised machine learning can help us reduce this computational burden.

MATHEMATICAL MODEL AND CHARACTERIZATION OF EMOTIONAL STATES IN BIPOLAR DISORDER

Jiménez Agudelo, Yury

CUNEF Universidad

Bipolar disorder is a mental illness caused by a lack of ability to manage changes in emotional states. Patients with bipolar disorder therefore suffer from states that are far from their own normality (euthymia) that lead to two critical states: depression and/or mania. One of the most frequent symptoms of any of these emotional states is sleep disturbance. Different studies have validated that changes in sleep patterns can be early indicators of a crisis in bipolar disorder. Therefore, a study of the sleep pattern as an early predictor of emotional crises is justifiable. This is a key aspect since it allows patients to have personalized medical treatment and prevent emotional crises. The crises in bipolar patients are triggered by endogenous and exogenous factors, and the latter are conditioned by specific environmental factors for each patient. As a first approximation in this direction, this study presents a mathematical model that characterizes the euthymic profile of a patient and with it the deviations that occur over time towards crisis states are detected. In the first instance, patients' data are collected to characterize their euthymic profile, and indices related to their emotional state are generated, such as quality of sleep, physical activity, and consumption of coffee, etc. Methods of dimensionality reduction and analysis of the correlation between variables and other methods based on Machine Learning (ML) are applied to classify and isolate the minimum set of variables. In the second instance, the use of portable sensors to collect data is analysed since some can be very intrusive and therefore rejected by patients. At this stage, a first approximation of a non-wearable sleep monitoring system that detects sleep time, the time a person remains lying down and the quality of sleep is proposed. The purpose of this study is to advance in the prediction of an emotional crisis based on sleep patterns and justifying that these records contain sufficient information to characterize the euthymic profile and its deviations, in the sense of the mathematical model proposed."

MACHINE LEARNING APPLIED TO CREDIT MODELS WITHOUT CLOSED FORM SOLUTION

Mahari, Thandiwe Irina
BBVA

In the financial modelling world there are a plethora of interesting models for which no analytical solution exists. When an analytical solution is not possible, a numerical solution must be used, for example a Monte Carlo simulation. Those kinds of calculations are very costly computationally speaking, which means that the family of models without analytical solution is rarely used in practice, regardless of other particular advantages.

In our particular test case we try to solve a problem related to credit modelization. It is difficult to find a credit model that generates a volatility as large as the one endured during the crisis period of 2009-2011. We have chosen a CIR model with stochastic volatility, a model with attractive properties but belonging to the family of models without closed form solution.

Furthermore, following the work of Savine and Huge (2020), we propose the replacement of the classical Monte Carlo by a neural network trained with stochastic sampling of valuation paths. In our presentation, we will discuss preliminary results that have been obtained, using the default probability by a CIR model with stochastic volatility. As a result, the new algorithm turns out to be simpler and noticeably more efficient.

MODEL-DRIVEN OR DATA-DRIVEN? PRICING AND PREDICTION IN MULTI-DEALER-TO-CLIENT PLATFORMS FOR EUROPEAN GOVERNMENT BONDS

Marín Martínez, Paloma
BBVA

As in many other areas, digitization has also gained great importance in financial markets, moving from voice to electronic channels. Multi-Dealer-to-Client (MD2C) platforms allow clients to simultaneously request several dealers for quotes. In these platforms, dealers compete for the same transactions, and they do not see the other dealers' prices. If a dealer is able to predict the probability that the client will accept a given price, she can use this information for pricing the operation. We have explored different approaches for this prediction, which can be model-driven or data-driven. On the one hand, we have used a modelization of the Request for Quote (RFQ) process (Fermanian, Guéant, & Pu, 2017). On the other hand, we have used machine learning techniques like logistic regression or neural networks. We discuss the advantages and disadvantages of both approaches. The research is based on a large proprietary database of RFQs about Italian and Spanish government bonds sent, through different MD2C platforms like Bloomberg, BondVision, and Tradeweb, to BBVA.

LEARNING A BATTERY OF COVID-19 MORTALITY PREDICTION MODELS BY MULTI-OBJECTIVE OPTIMIZATION

Martínez García, Mario
Basque Center for Applied Mathematics

The COVID-19 pandemic is continuously evolving with drastically changing epidemiological situations which are approached with different decisions: from the reduction of fatalities to even the selection of patients with the highest probability of survival in critical clinical situations. Motivated by this, a battery of mortality prediction models with different performances has been developed to assist physicians and hospital managers. Logistic regression, one of the most popular classifiers within the clinical field, has been chosen as the basis for the generation of our models. Whilst a standard logistic regression only learns a single model focusing on improving accuracy, we propose to extend the possibilities of logistic regression by focusing on sensitivity and specificity. Hence, the log-likelihood function, used to calculate the coefficients in the logistic model, is split into two objective functions: one representing the survivors and the other for the deceased class. A multi-objective optimization process is undertaken on both functions in order to find the Pareto set, composed of models not improved by another model in both objective functions simultaneously. The individual optimization of either sensitivity (deceased patients) or specificity (survivors) criteria may be conflicting objectives because the improvement of one can imply the worsening of the other. Nonetheless, this conflict guarantees the output of a battery of diverse prediction models. Furthermore, a specific methodology for the evaluation of the Pareto models is proposed. As a result, a battery of COVID-19 mortality prediction models is obtained to assist physicians in decision-making for specific epidemiological situations.

DESIGNING EXPERIMENTS FOR ESTIMATING AN APPROPRIATE OUTLET SIZE FOR A SILO TYPE PROBLEM

Moler Cuiral, José Antonio
Universidad Pública de Navarra

Jam formation is a problem that may occur when granular material is discharged by gravity from a silo. The estimation of the minimum outlet size which guarantees that the time to the next jamming event is long enough can be crucial in the

industry. The time is modeled by an exponential distribution with two unknown parameters, and this goal translates to precise estimation of a non-linear transformation of the parameters.We obtain c-optimum experimental designs with that purpose, applying

the graphic Elfving method. Because the optimal experimental designs depend on the nominal values of the parameters, we conduct a sensitivity analysis on our dataset.

Finally, a simulation study checks the performance of the approximations, first with the Fisher Information matrix, then with the linearization of the function to be estimated.

The results are useful for experimenting in a laboratory and translating then the results to a real scenario. From the application, we develop a general methodology for estimating a one-dimensional transformation of the parameters of a nonlinear model.

Keywords and phrases: bulk solid storage, jam formation, non-linear heteroscedastic model, optimal design of experiments.

USING HIGH-RESOLUTION VOLTAGE MAPS TO PREDICT "REDO" IN THE TREATMENT OF ATRIAL FIBRILLATION (AF)

Jean Bragard [1], *Leire Moriones* [1], *Blas Echebarria* [2], *Susana Ravassa*[3]

[1] *School of Sciences, Universidad de Navarra, Pamplona, Spain*

[2] *Polytechnic University of Catalonia, Barcelona, Spain*

[3] *CIMA, Universidad de Navarra, Pamplona, Spain.*[1]

Aims: High-resolution voltage maps (HRVM) are used to predict the post-ablation recurrence of AF. This study aims to assess the statistical power of electrical biomarkers extracted from the HRVM. This paper is a follow-up

from a previous analysis. Now the number of patients in the cohort has been augmented from 98 to 139. Atrial fibrillation recurrence (AFR) is related tolower mean voltage of the patient left atrium.

Methods: With the same catheter used in the ablation procedure an acqui-sition of HRVM was performed on the left atrium The resolution of the maps are the solution for AF patients in 40-70% cases. Bipolar voltage map is estimated with two electrical biomarkers and one geometrical characteristic (Area). Supervised classifier (from Matlab Machine Learning Toolbox) is used, specifically, the logistic regression and the coarse tree classifiers.

Results: AUC, accuracy and confusion matrices were compared between the two classifiers. For the cohort of 98 patients, weighted KNN model shows an accuracy of 80.6% and a ACU=0.81; logistic regression accuracy`s is 76.5% and a ACU=0.74; coarse tree`s accuracy is 70.4% and a ACU=0.63. The cohort of 139 patients has a logistic regression accuracy of 77.0% and a AUC=0.74; coarse tree`s accuracy is 77.0% and a AUC=0.57.

Conclusions: Surprisingly, we have noted that the classification has slightly worsen with respect to the previous paper when the cohort was formed with 98 patients. Several explanations are plausible. The most obvious is that the HRVM are not enough to predict redo with very high accuracy.

A more comprehensive classifiers with combination of clinical, demographical and comorbidities should presumably improve the prediction of future redo procedure for a patient.

ALGORITHMIC MARKET-MAKING AND INFORMATION ASYMMETRY

Ojeda Trejo, Christian

BBVA

One of the main roles of dealer-banks / market-makers in financial markets is the provision of liquidity to other market participants. Market-makers stand ready to quote prices at which they are willing to buy or sell financial instruments, trying to make a profit on the bid-ask spread. As such, they have to deal with the problem of quoting optimal prices that compensate them for the underlying risks associated with liquidity provision, namely price risk in their inventory holdings and information asymmetry. Motivated by the context that dealer-banks face in multi-dealer to client platforms based on the request-for-quote (RfQs) protocol like Tradeweb and Bloomberg, in this work (i) we present a general optimal market-making modeling framework that extends the seminal model of Avellaneda-Stoikov (2008) to include specific features of quote-driven markets, where client identity and volumes requested are known for ongoing requests but uncertain for future ones (ii) we also obtain new closed-form approximations for the optimal quotes in this approach (iii) we extend the model to incorporate a continuous-time version of the Glosten & Milgrom model (1985) for optimal quotes with heterogeneously informed traders and (iv) we present an algorithm based in deep learning to numerically approximate optimal quotes under information asymmetry.

YOU SAY IT AND IT DOES IT TO YOU ARTIFICIAL INTELLIGENCE ALREADY ALLOWS PROGRAMMING TO BE DEMOCRATIZED TO ANY PROFILE

Javier Porras Castaño, Javier
Unicaja Banco - Doctoral student in Artificial Intelligence

Gpt-31 (Codex2 , Github Copilot3), AlphaCode4 , Megatron-Turing5 , among others, are examples of a new powerful capacity of Artificial Intelligence (AI) that will represent a turning point in the "low code" trend, allowing programming for you just by saying it with your own form of expression and without having any idea of programming (without being a technical profile).

This new AI prowess is rooted in a new neural network architecture, called Transformers6 : it is revolutionizing natural language processing to automate its understanding. This is achieving new and attractive use cases for the business environment, such as democratizing programming to any profile.

Although it is only in its initial phase or taking off, it could already begin to add value to companies, mainly to accelerate software development at two levels:

1: If you are a programmer

> It will help you get your work done faster: the programmer only has to tell it what he wants, in natural language, and it implements it as is.
>
> Perhaps it is no longer necessary to master a specific programming language (Python, Java, Node, R or C++) because you will only need to clearly express what you want to do and the AI takes care of implementing it in any language.

2: If you are not a programmer

> It will allow the development of software (mockups, prototypes, proofs of concept or minimum viable products) minimizing dependence on a company's technology department. You only need to tell it what you want to do, with your form of expression (text or voice) and you will have it done.

For the first time, we will be able to assign resources (people) to any software project regardless of its technical characteristics.

References:
1: https://openai.com/blog/gpt-3-apps/
2: https://openai.com/blog/openai-codex/
3: https://copilot.github.com
4: https://alphacode.deepmind.com
5: https://arxiv.org/abs/2201.11990
6: https://arxiv.org/abs/1706.03762

ON THE COMPUTATION OF TRIANGULAR NORMS IN CHAINS

Francisco Javier Talavera Andújar y Jorge Elorza Barbajero
Universidad de Navarra

In the field of fuzzy logic, much effort has been devoted to generalise the operators of classical set theory (union, intersection, etc.). In the case of the intersection, these operators are known as triangular norms (t-norms for short). Originally, they were defined in the unit square $[0,1]^2$ but, in 1994, De Cooman and Kerre extended them to bounded lattices.

Chains are the simplest lattices as all their elements are comparable. Moreover, most applications make use of chains. That is why we will focus on them, although the results can be adapted to any kind of lattice. We will provide a recursive algorithm to find all t-norms in a chain of n+1 elements from the t-norms in a chain of n elements in order to reduce computational costs.

CHARACTERIZING THE INTERPLAY BETWEEN BIOLOGICAL FACTORS AND CELL GROWTH IN UNPERTURBED TUMOR GROWTH DYNAMICS

Aymara Sancho-Araiz (1,2), Zinnia P Parra-Guillen (1,2), Víctor Mangas-Sanjuan (3,4), Iñaki F. Trocóniz (1,2)

(1) Department of Pharmaceutical Technology and Chemistry, School of Pharmacy and Nutrition, University of Navarra, Pamplona, Spain. (2) Navarra Institute for Health Research (IdiSNA), Pamplona, Spain. (3) Department of Pharmacy Technology and Parasitology, Faculty of Pharmacy, University of Valencia, Valencia, Spain. (4) Interuniversity Institute of Recognition Research Molecular and Technological Development. Valencia, Spain

Objectives:

Mathematical modeling of unperturbed and perturbed tumor growth dynamics (TGD) in preclinical experiments provides an opportunity to establish translational frameworks [1,2]. Most of the commonly used models describe natural growth with a basic function: linear, exponential, Gompertz. More complex models aimed to include tumor heterogeneity or biologic processes have also been described [1]. Despite this, tumor growth curves of these models follow a monotonic increase, and although they tend to capture individual TGD and variability in the data reasonable well, systematic model misspecifications can be identified. This represents an opportunity to investigate possible underlying mechanisms controlling tumor growth dynamics. The overall goal of this work is to develop a mathematical model to describe tumor growththat can be systematically applied during the preclinical evaluation of new drug candidates.

Methods: Tumor volumes (TV) of 12 different cell lines from 6 tumor types (breast, leukemia, lung, lymphoma, melanoma, and pancreas) were available for the analysis. TVs from the different cell lines were analyzed using NONMEM 7.4. Different models ranging from more empirical to more mechanistic were explored. The model building was performed sequentially. First, the unperturbed tumor growth was characterized using previous models [2], and then the new structure was established. Numerical and graphical metrics, including residual-based diagnostics (weighted residuals and autocorrelation plots) and visual predictive checks (VPCs), were explored and compared for model selection and evaluation.

Results:

All the data were analyzed simultaneously through a joint modeling exercise, using the type of cell line as a categorical covariate. Gompertz TGD model, in which the relative

tumor growth rate decreases until reaching its maximum carrying capacity, provided a good description of the data and was used as a core structure. The estimate of the first-order growth rate constant (kge), ranges from 0.0192 – 0.0951 $days^{-1}$, with an inter-animal variability of 14.6%. With regard to the initial tumor size (TV_0) and the maximum carrying capacity (Tmax), the estimates ranged from 10.2-56 mm^3 and 228 – 10000 mm^3, respectively. Despite VPCs and classical basic goodness of fit indicated and adequate model performance, a systematic missespecification was detected when exploring weighted residuals versus time and autocorrelation plots. From the different structures evaluated to describe the growth dynamics observed in tumor growth over time, the final model developed included an increasing tumor growing capacity dependent on the amount of nutrients and vasculature. In this regard, under restrictive conditions, i.e. the tumor size is a 10% lower than the growing capacity value, cancer cells trigger a signal able to initiate angiogenesis in order to further enable tumor regrowth. The new model significantly improved overall model performance, especially showing an improvement in the weighted residuals versus time and the autocorrelation plots.

Conclusions:

Systematical model misspecifications have been identified when using standard models to describe xenograft tumor growth dynamics from different tumor types and cell lines in preclinical arena. This work presents a new semi-mechanistic model capable of describing the non-monotonic growth and the interactions between tumor growth, angiogenesis, and nutrient supply. This framework constitutes a valuable tool to explore different mechanisms of action, thus supporting the rational design and selection of drug scenarios in monotherapy or combination during preclinical drug development.

References:

[1] Yin, A.; Moes, D.J.A.R.; van Hasselt, J.G.C.; Swen, J.J.; Guchelaar, H.J.; A, Y.; DJAR, M.; JGC, van H.; JJ, S.; HJ, G. A Review of Mathematical Models for Tumor Dynamics and Treatment Resistance Evolution of Solid Tumors. *CPT pharmacometrics Syst. Pharmacol.* **2019**, *8*, 720–737, doi:10.1002/PSP4.12450.

[2] Benzekry, S.; Lamont, C.; Beheshti, A.; Tracz, A.; Ebos, J.M.L.L.; Hlatky, L.; Hahnfeldt, P. Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth. *PLoS Comput. Biol.* **2014**, *10*, e1003800, doi:10.1371/journal.pcbi.1003800

AI-DRIVEN ORGANIZATION MATURITY: A CONCEPTUAL FRAMEWORK

Yablonsky, Sergey

Universidad de Jaén

Purpose – To be more effective, artificial intelligence (AI) requires a broad overall view of the design and transformation of enterprise architecture and capabilities. Maturity models (MMs) are the recognized tools to identify strengths and weaknesses of certain domains of an organization. They consist of multiple, archetypal levels of maturity of a certain domain and can be used for organizational assessment and development. In the case of AI, quite a few numbers of MMs have been proposed. Generally, the links between AI technology, AI usage and organizational performance stay unclear. To address these gaps, this paper aims to introduce the complete details of the AI maturity model (AIMM) for AI-driven companies. The associated AI-Driven Enterprise Maturity framework proposed here can help to achieve most of the AI-driven companies' objectives.

Design/methodology/approach – Qualitative research is performed in two stages. In the first stage, a review of the existing literature is performed to identify the types, barriers, drivers, challenges and opportunities of MMs in AI, Advanced Analytics and Big Data domains. In the second stage, a research framework is proposed to align company value chain with AI technologies and levels of the enterprise maturity.

Findings – The paper proposes a new five level AI-Driven Enterprise Maturity framework by constructing a formal organizational value chain taxonomy model that explains a vast group of MM phenomena related with the AI-Driven Enterprises. In addition, this study proposes a clear and precise description and structuring of the information in the digital platform, AI, Advanced Analytics and Big Data domains. The AI-Driven Enterprise Maturity framework assists in identification, creation, assessment, and disclosure research of AI-driven organizations.

Research limitations/implications – This research is focused on the basic dimensions of AI value chain.