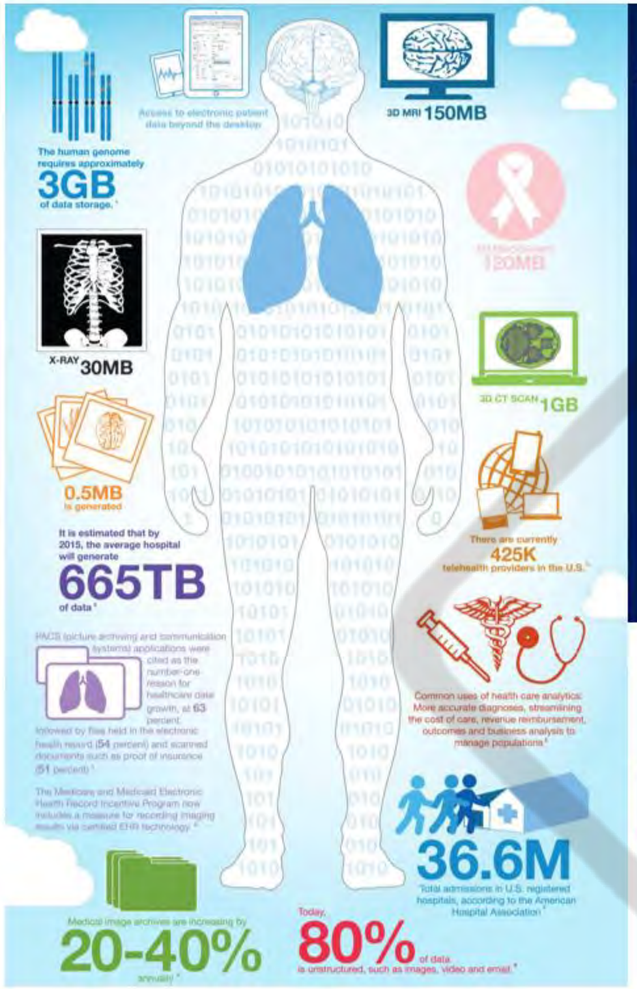
eXplainable Artificial Intelligence for the identification of novel biomarkers of disease

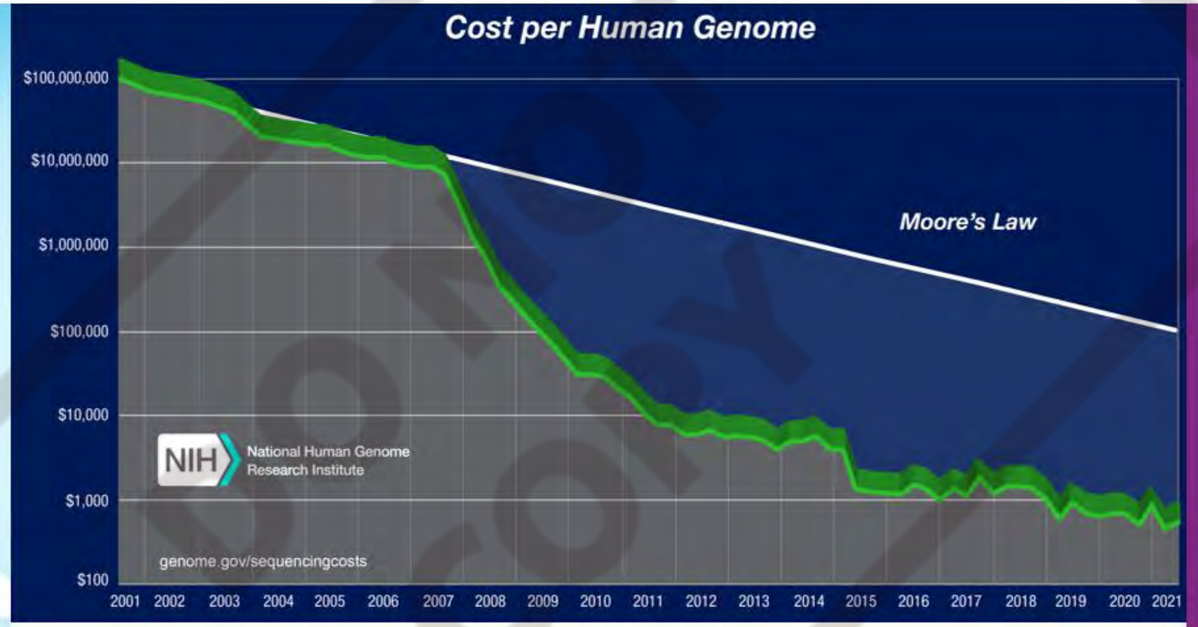
Augusto Anguita-Ruiz, Postdoctoral Research Fellow at ISGlobal

Barcelona Biomedical Research Park (PRBB) , Doctor Aiguader, 88, 08003 Barcelona, Spain

Dec, 9th 2024

Information sources





"Human body is an unlimited sourde of data "

Business Analytics

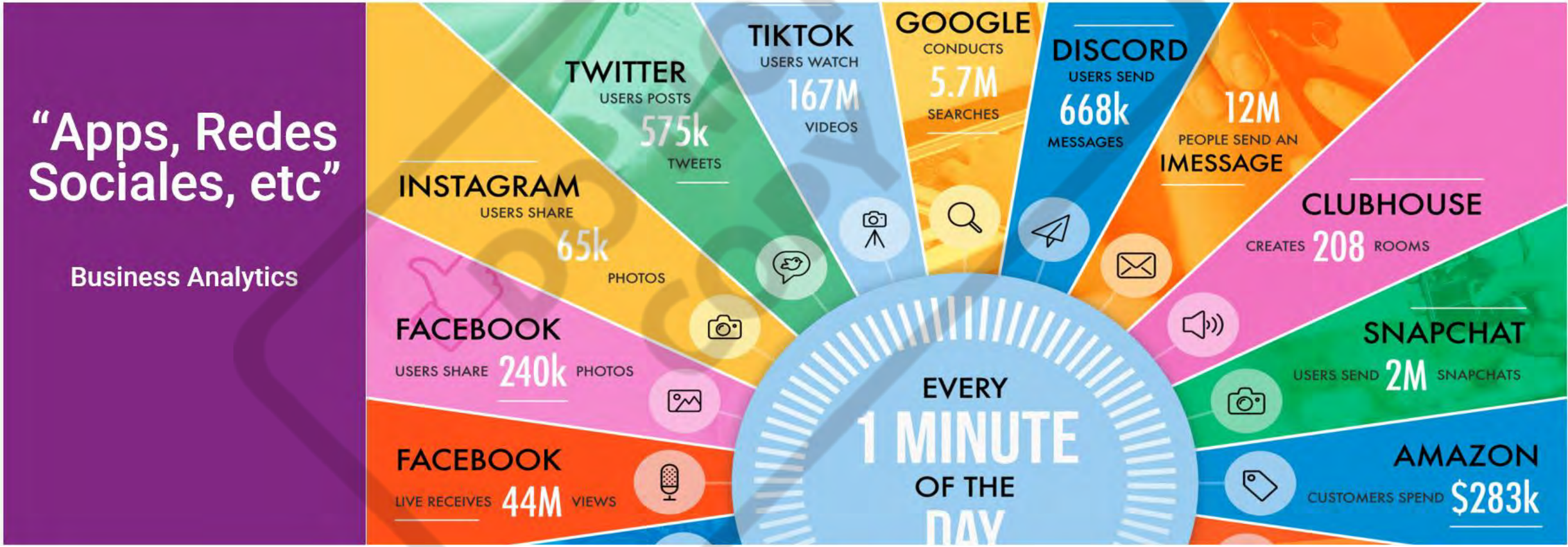
New biotechnologies and cost reduction

Lower costs and increased technology performance are driving a veritable tsunami of data that is enabling huge advances in these disciplines in just a few years.

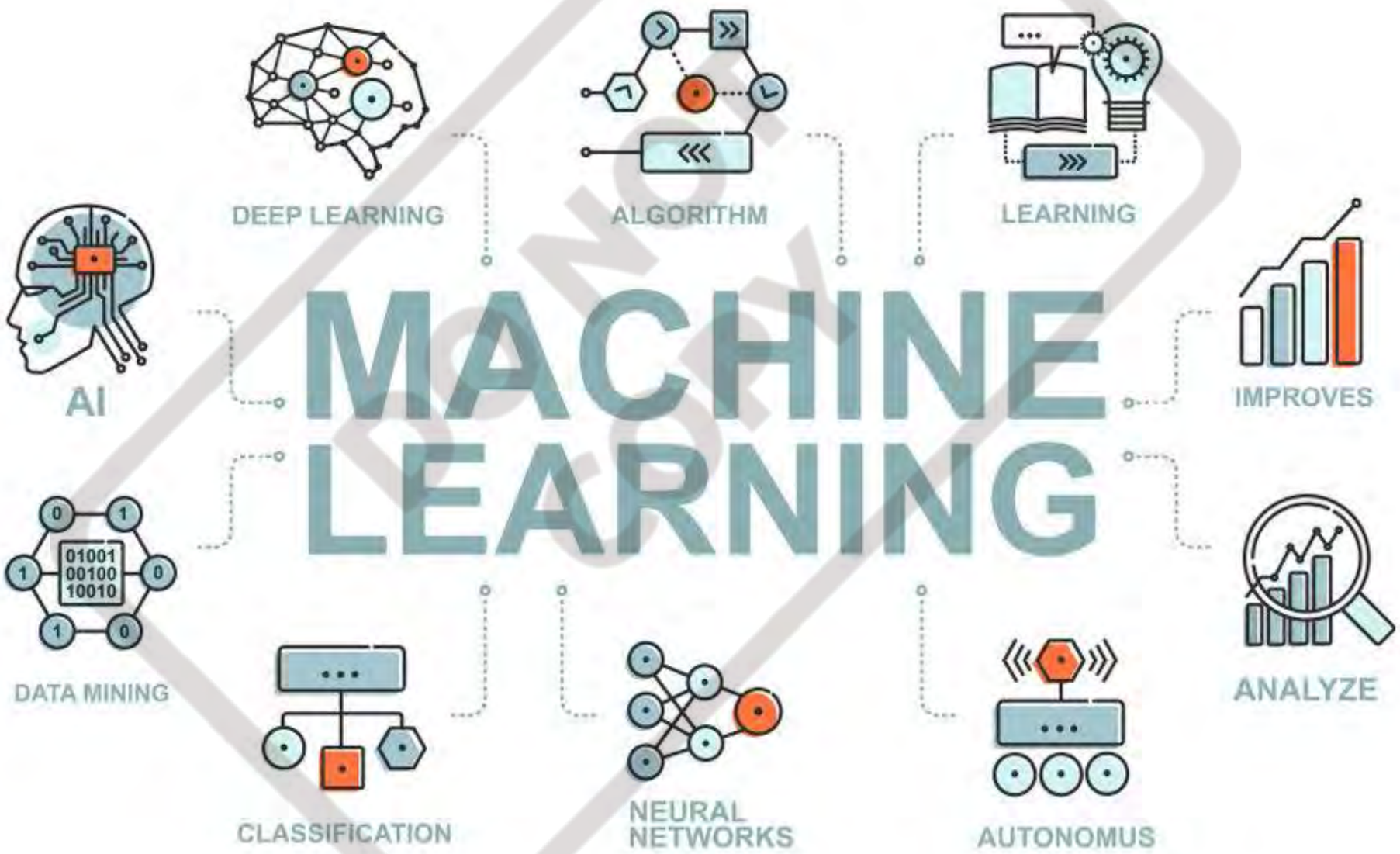
Information sources

Internet and social media

An unfathomable amount of digital activity is occurring at any given moment. This ongoing explosion in activity is the aggregate output 4.5 billion internet users today, a number that's projected to increase even further in coming years



AI & Machine learning



AI & Machine learning

- The goal of Machine Learning is creating statistical **models**
- **Represent** simplified real-world process with statistics
- Mathematical **relationships** between variables
- Based on statistical assumptions and **historical data**
- Implies an evaluation **metric**

AI & Machine learning

DESCRIPTIVE
To understand **what** happened

PREDICTIVE
To forecast what **will** happen

DIAGNOSTIC
To determine **why did** it happened

PRESCRIPTVE
To establish **how** can we **make** it happen

AI & Machine learning

DESCRIPTIVE
To understand **what** happened

PREDICTIVE
To forecast what **will** happen

DIAGNOSTIC
To determine **why did** it happened

PRESCRIPTVE
To establish **how** can we **make** it happen

Machine Learning

Unsupervised Learning: Association Rules

DISCOVERY OF ASSOCIATIONS

Discovery of rules or patterns which are used to represent dependencies between data/variables of a data bases.

Machine Learning

Unsupervised Learning: Association Rules

DISCOVERY OF ASSOCIATIONS

Discovery of rules or patterns which are used to represent dependencies between data/variables of a data bases.

Machine Learning

Unsupervised Learning: Association Rules

DISCOVERY OF ASSOCIATIONS

Discovery of rules or patterns which are used to represent dependencies between data/variables of a data bases.

IF buy **beer**, THEN buy **diapers** — 75%

IF buy **diapers**, THEN buy **beer** — 100%

Example of using association rules with omics

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research

Augusto Anguita-Ruiz[1,2,3]*, Alberto Segura-Delgado[4], Rafael Alcalá[4], Concepción M. Aguilera[1,2,3], Jesús Alcalá-Fdez[4]

1 Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Granada, Spain, **2** Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain, **3** CIBEROBN (Physiopathology of Obesity and Nutrition), Instituto de Salud Carlos III (ISCIII), Madrid, Spain, **4** Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

* augustoanguitaruiz@gmail.com

SEQUENTIAL RULE MINING ALGORITHM

CMRules algorithm for mining sequential rules from temporal gene expression data:

Temporal gene networks

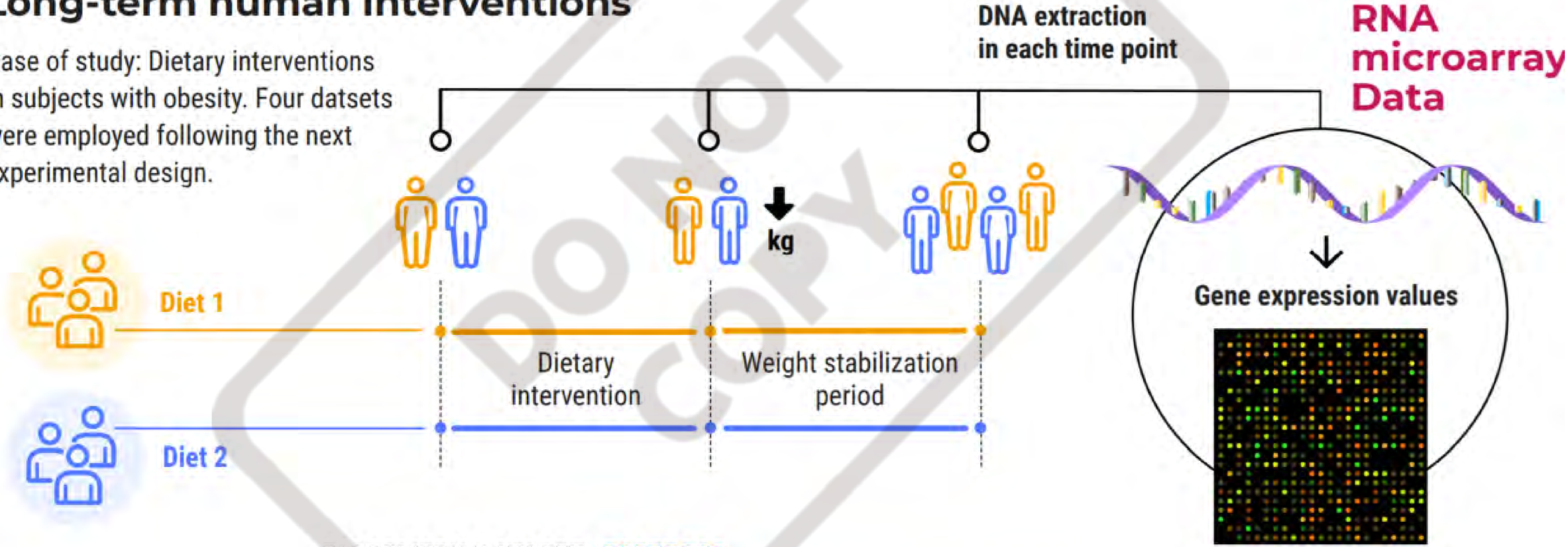
[gene A ↑ gene B ↓] ↓ time delay

[gene C ↑ gene D ↑ gene E ↑]

Example of using association rules with omics

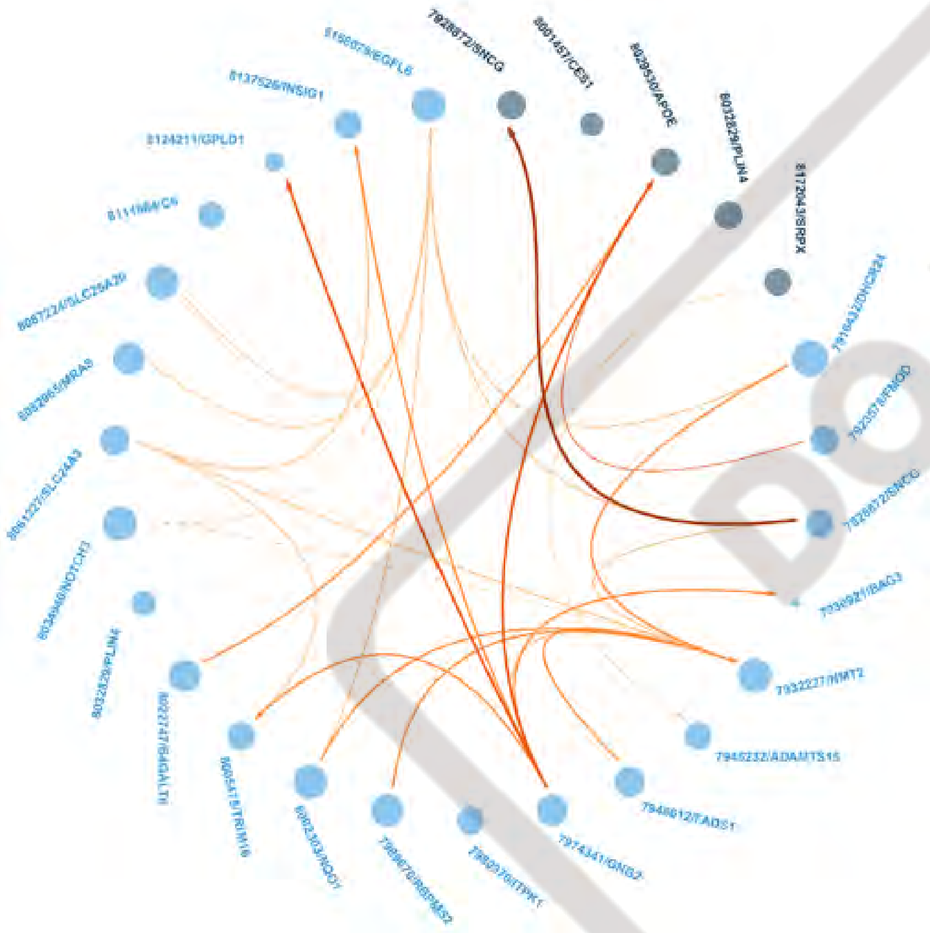
Long-term human interventions

Case of study: Dietary interventions in subjects with obesity. Four datsets were employed following the next experimental design.



DISCOVERY DATASET: GSE77962
VALIDATION DATASET: GSE70529, GSE35411 and GSE103766

Example of using association rules with omics



- ✓ The proposed method was **validated in six datasets** from obesity research (consisting of low-calorie diets interventions), where it was able **to extract meaningful gene-gene temporal interactions** with relevance in the etiology of the disease
- ✓ The application of such pipeline to other type of human temporal gene profiles would **greatly expand our knowledge for complex biological processes**, with a special interest for drug clinical trials, in which identified gene-gene regulatory interactions could reveal **new therapeutic targets**

AI & Machine learning

DESCRIPTIVE
To understand **what** happened

DIAGNOSTIC
To determine **why did** it happened

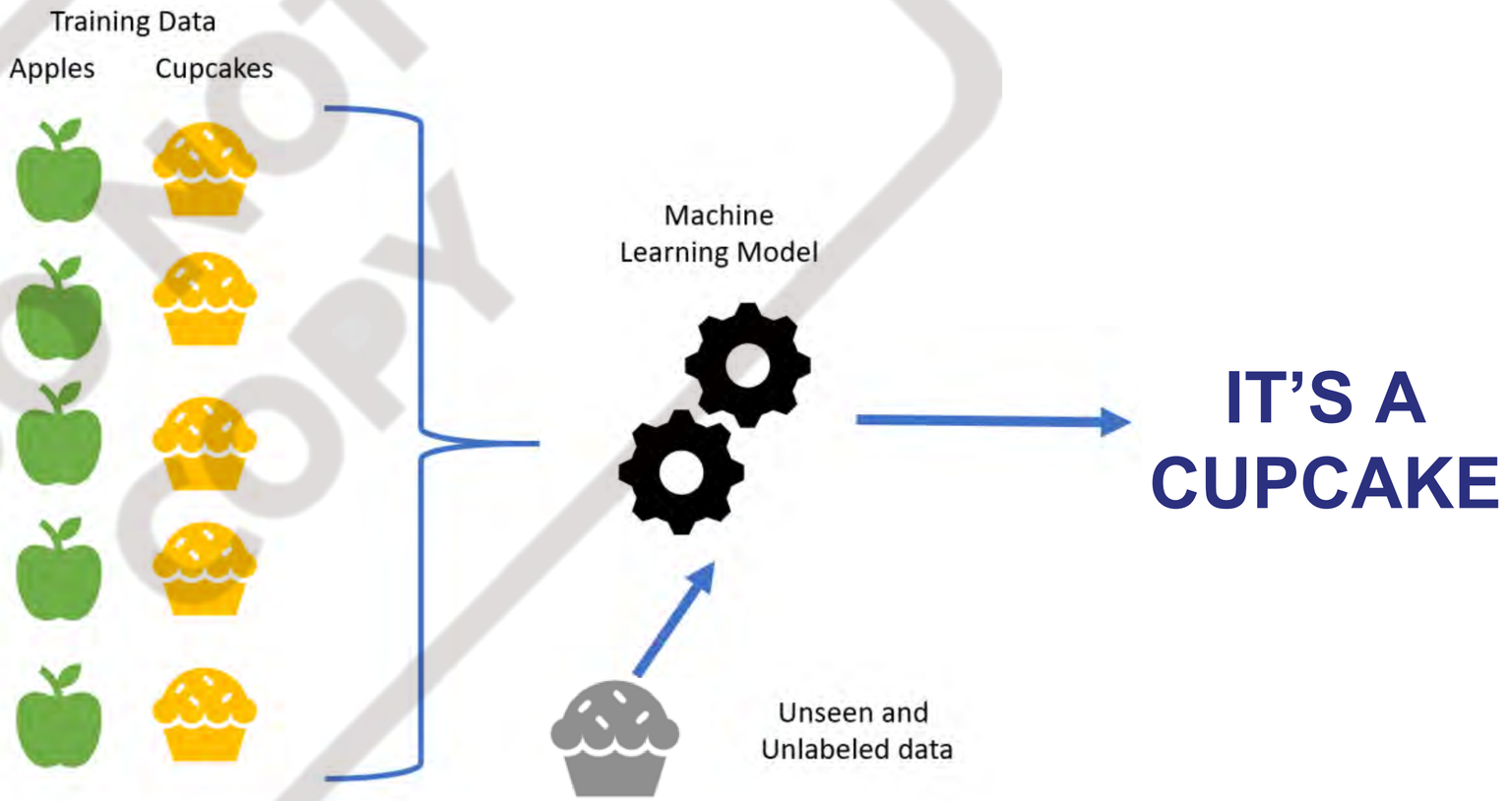
PREDICTIVE
To forecast what **will** happen

PRESCRIPTVE
To establish **how** can we **make** it happen

Machine Learning
Supervised Learning: Classification

CLASSIFICATION

The goal is to build a predictive model (classifier) using labeled training data. The labels can have binary or categorical values



Example of using XAI with predictive purposes

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression

Alberto Ramírez-Mena[a], Eduardo Andrés-León[b], Maria Jesus Alvarez-Cubero[a,c], Augusto Anguita-Ruiz[d], Luis Javier Martinez-Gonzalez[a,*], Jesus Alcala-Fdez[e]

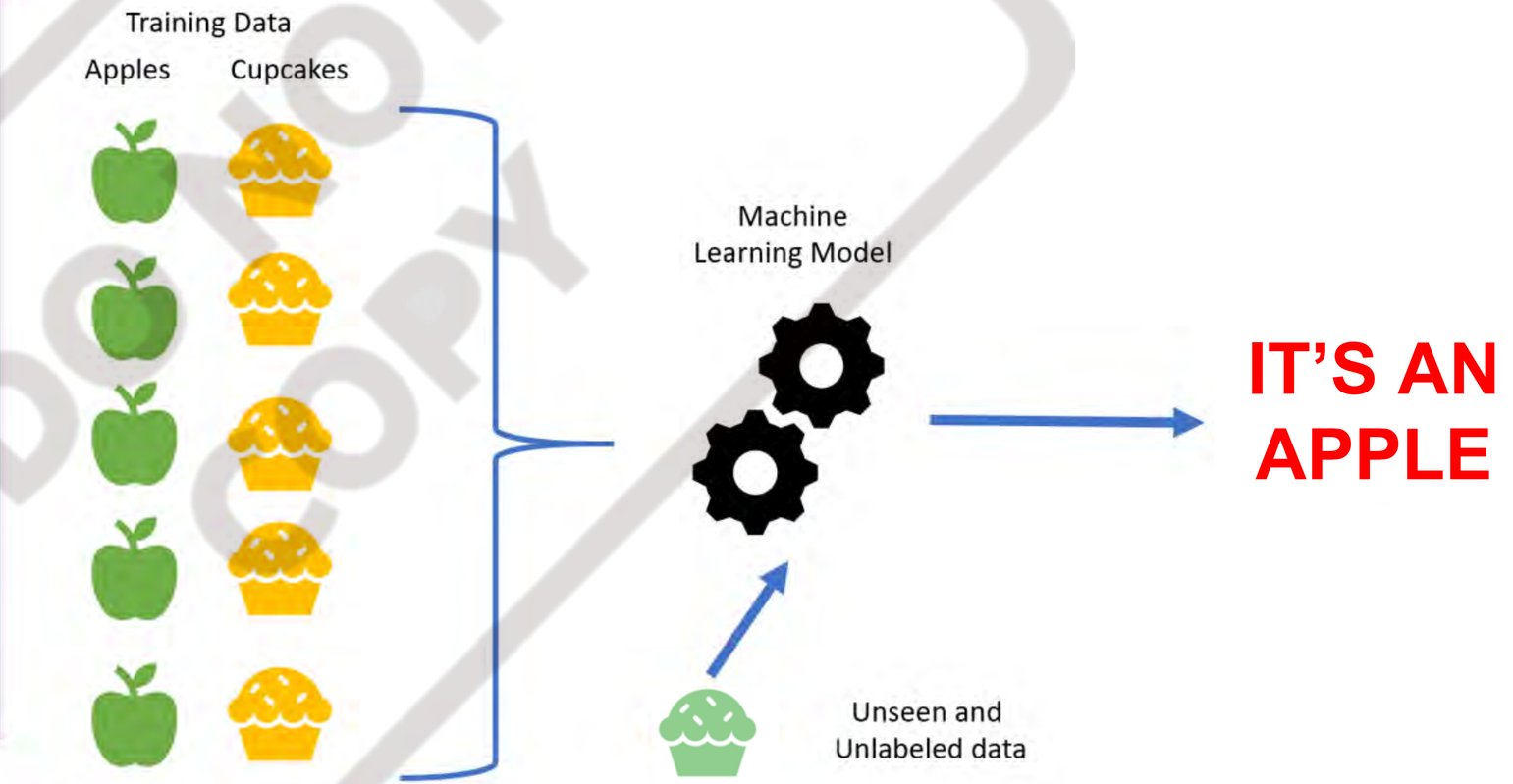
[a] GENYO, Centre for Genomics and Oncological Research: Pfizer -University of Granada – Andalusian Regional Government, Granada, 18016, Spain
[b] Institute of Parasitology and Biomedicine "López-Neyra" (IPBLN), Spanish National Research Council (CSIC), Granada, 18016, Spain
[c] Department of Biochemistry and Molecular Biology III and Immunology, University of Granada, Granada, 18071, Spain
[d] Barcelona Institute for Global Health, ISGlobal, Barcelona, 08003, Spain
[e] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain

Machine Learning

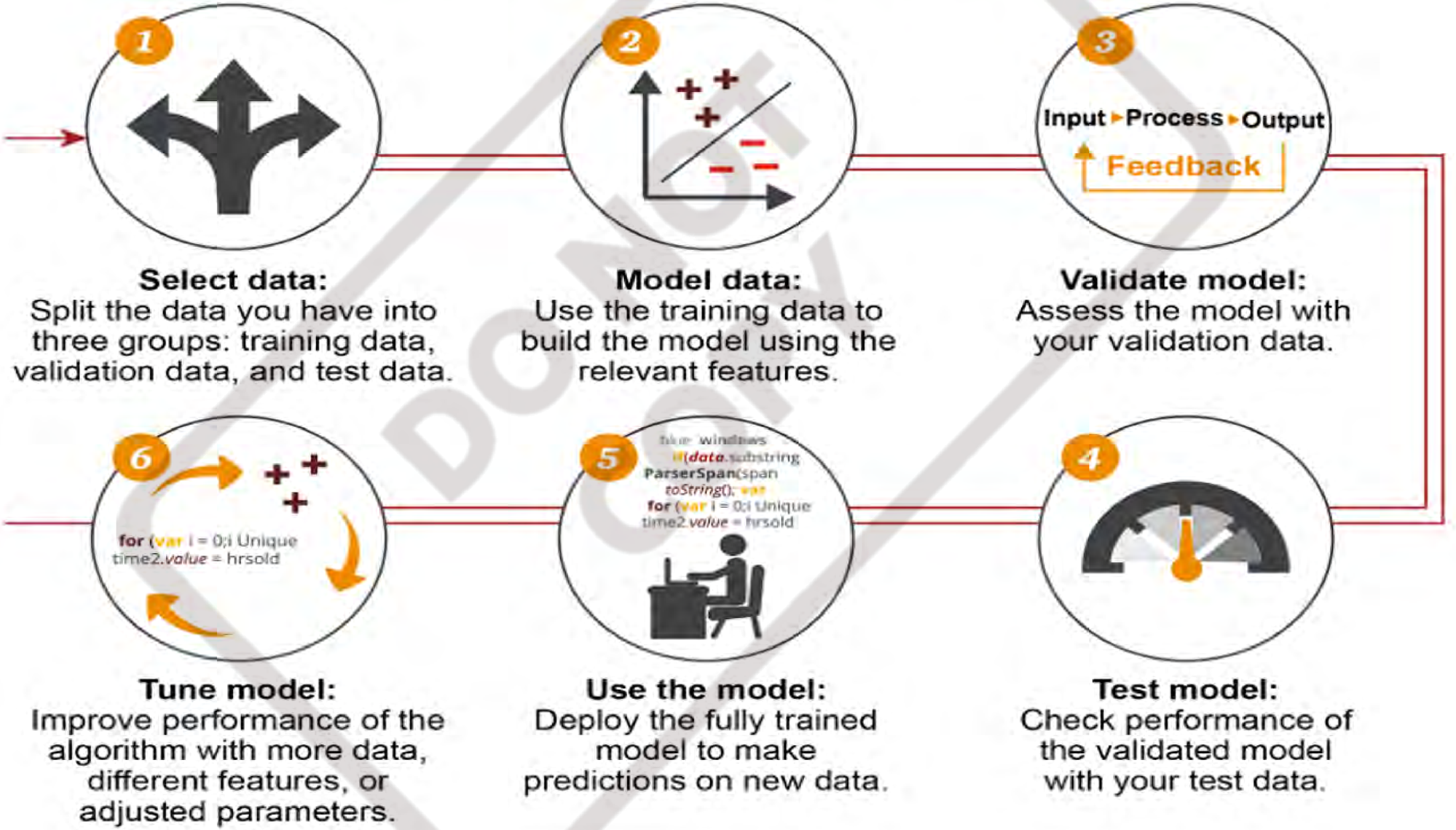
Supervised Learning: Classification

CLASSIFICATION

The goal is to build a predictive model (classifier) using labeled training data. The labels can have binary or categorical values



How Machine Learning Works



Select data:
Split the data you have into three groups: training data, validation data, and test data.

Model data:
Use the training data to build the model using the relevant features.

Validate model:
Assess the model with your validation data.

Tune model:
Improve performance of the algorithm with more data, different features, or adjusted parameters.

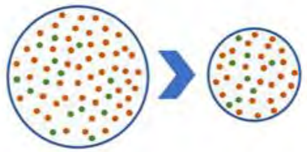
Use the model:
Deploy the fully trained model to make predictions on new data.

Test model:
Check performance of the validated model with your test data.

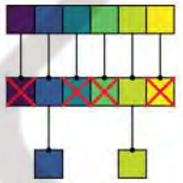
Data Preprocessing

Real-life data is “imperfect”

Preparation is done to prevent: errors, incorrect results, biasing algorithms



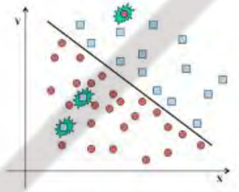
Instance Selection



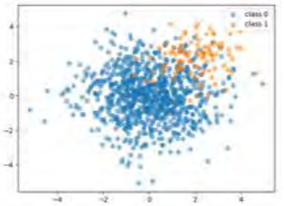
Features Selection

Data Transformations

Discretization



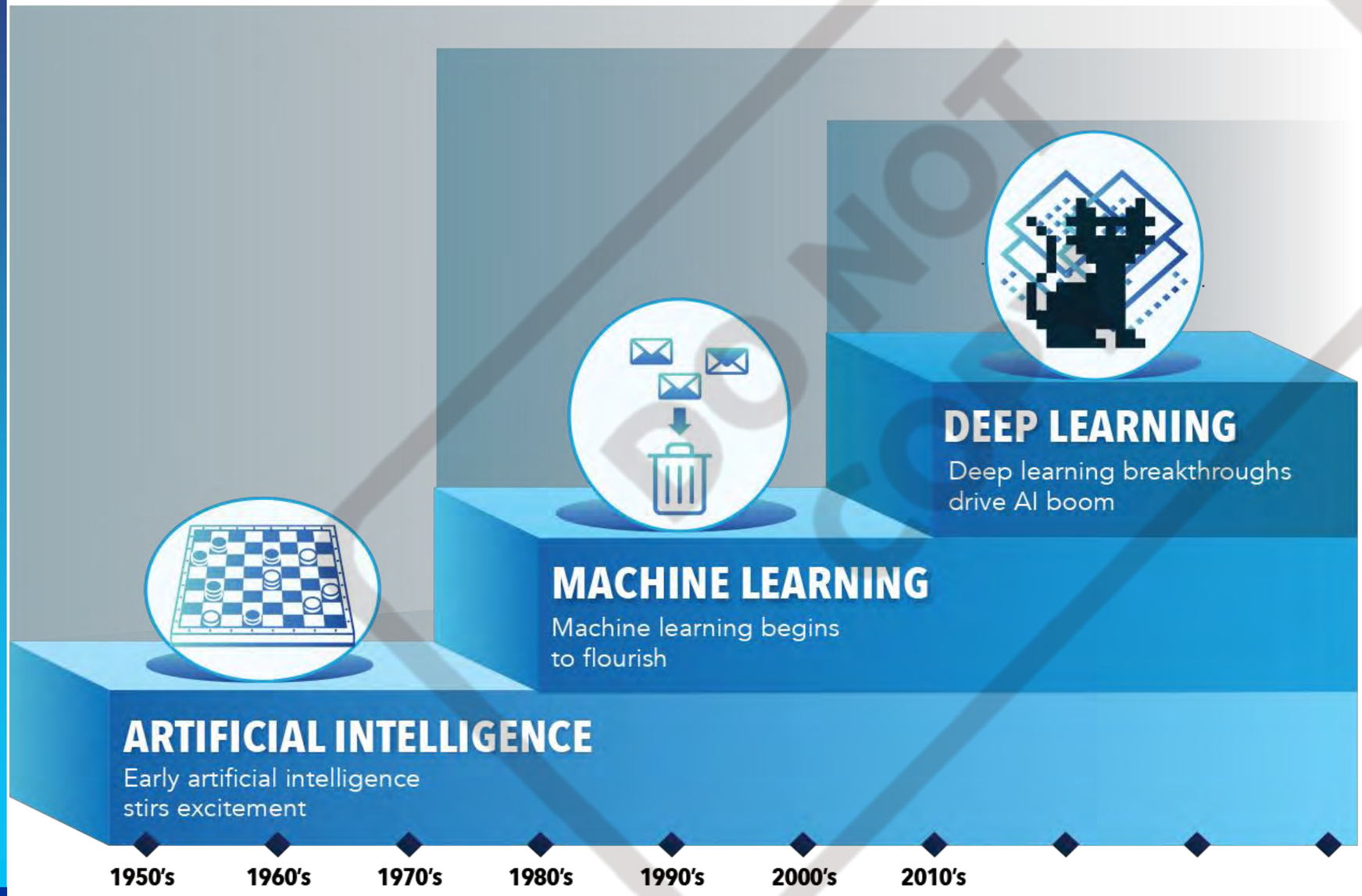
Noise Data



Imbalanced Classification

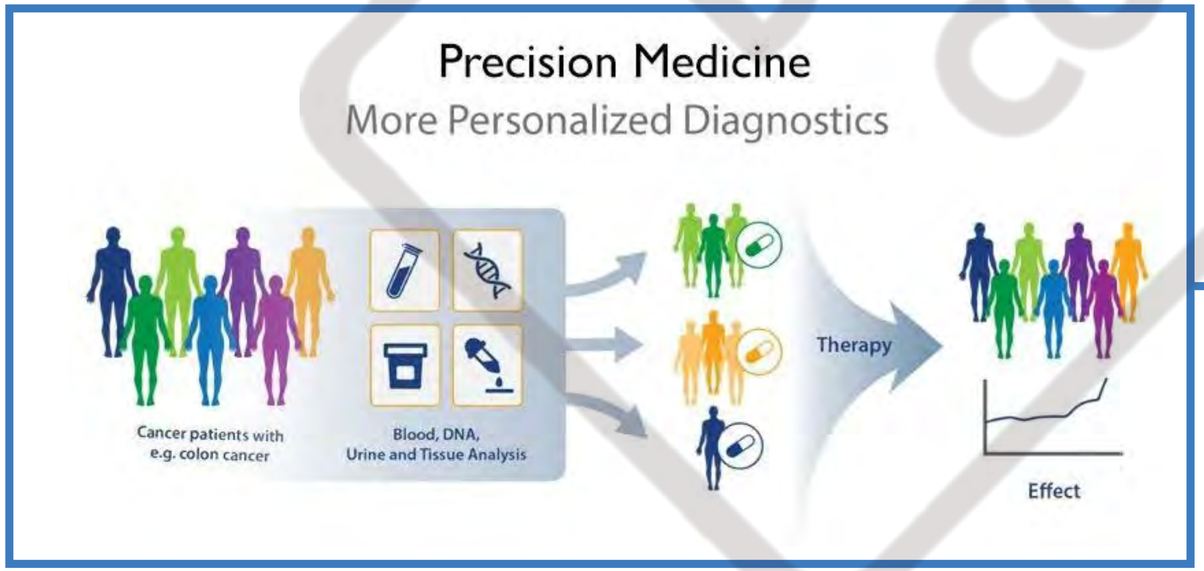
Dataset Shift

AI FOR PREDICTION



Unprecedented ability to make accurate predictions

PRECISION MEDICINE



1 RISK ASSESSMENT
Genetic testing to reveal predisposition to disease

2 PREVENTION
Behavior/lifestyle treatment intervention to prevent disease[2]

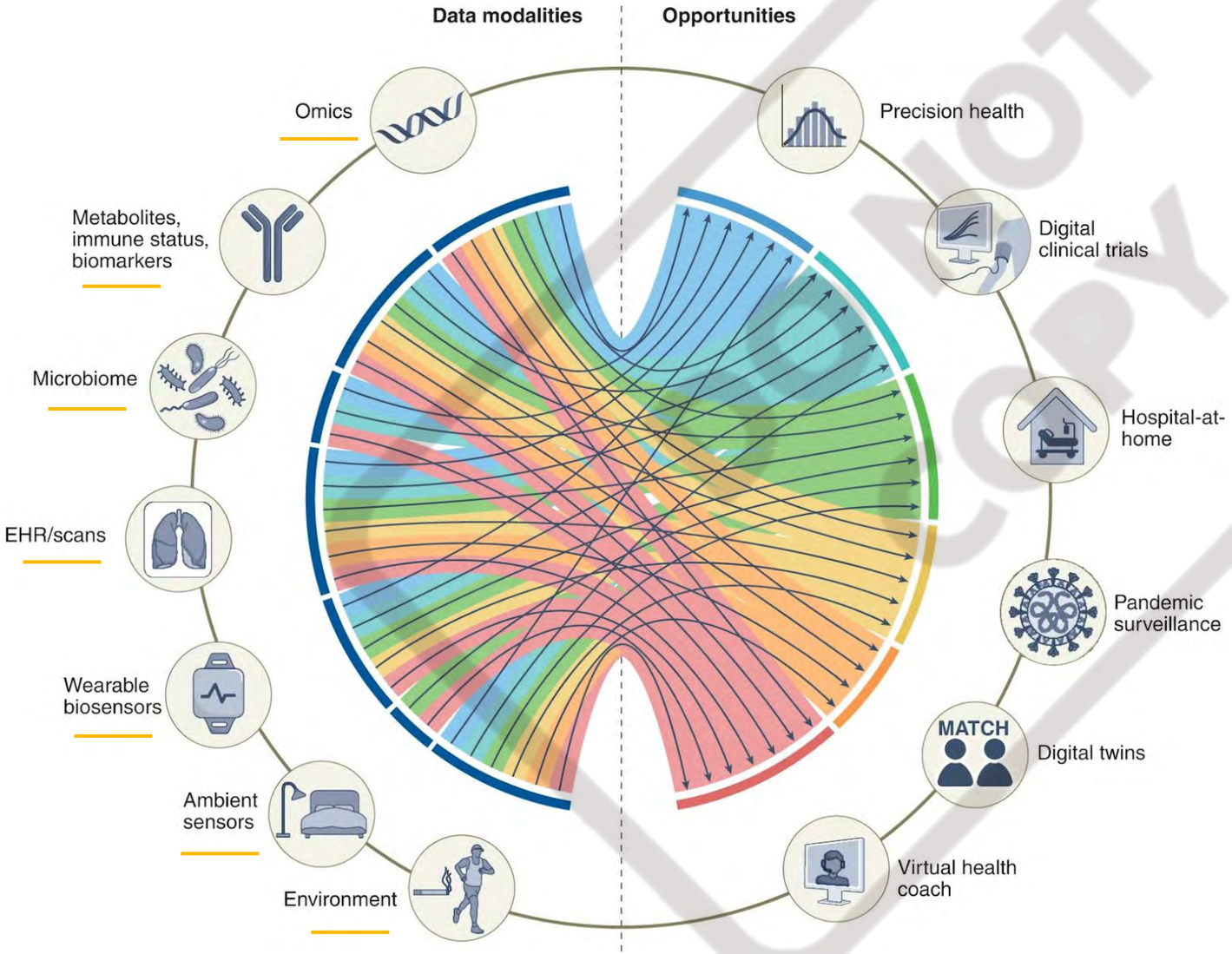
3 DETECTION
Early detection of disease at the molecular level when treat-ment can be most effective[3,4]

4 DIAGNOSIS
Accurate diagnosis to enable effective and individualized treatment strategy

5 TREATMENT
Targeted treatments to improve outcomes and reduce side effects

6 MANAGEMENT
Active monitoring of treatment response and disease progression

DATA INTEGRATION TO POWER PRECISION MEDICINE

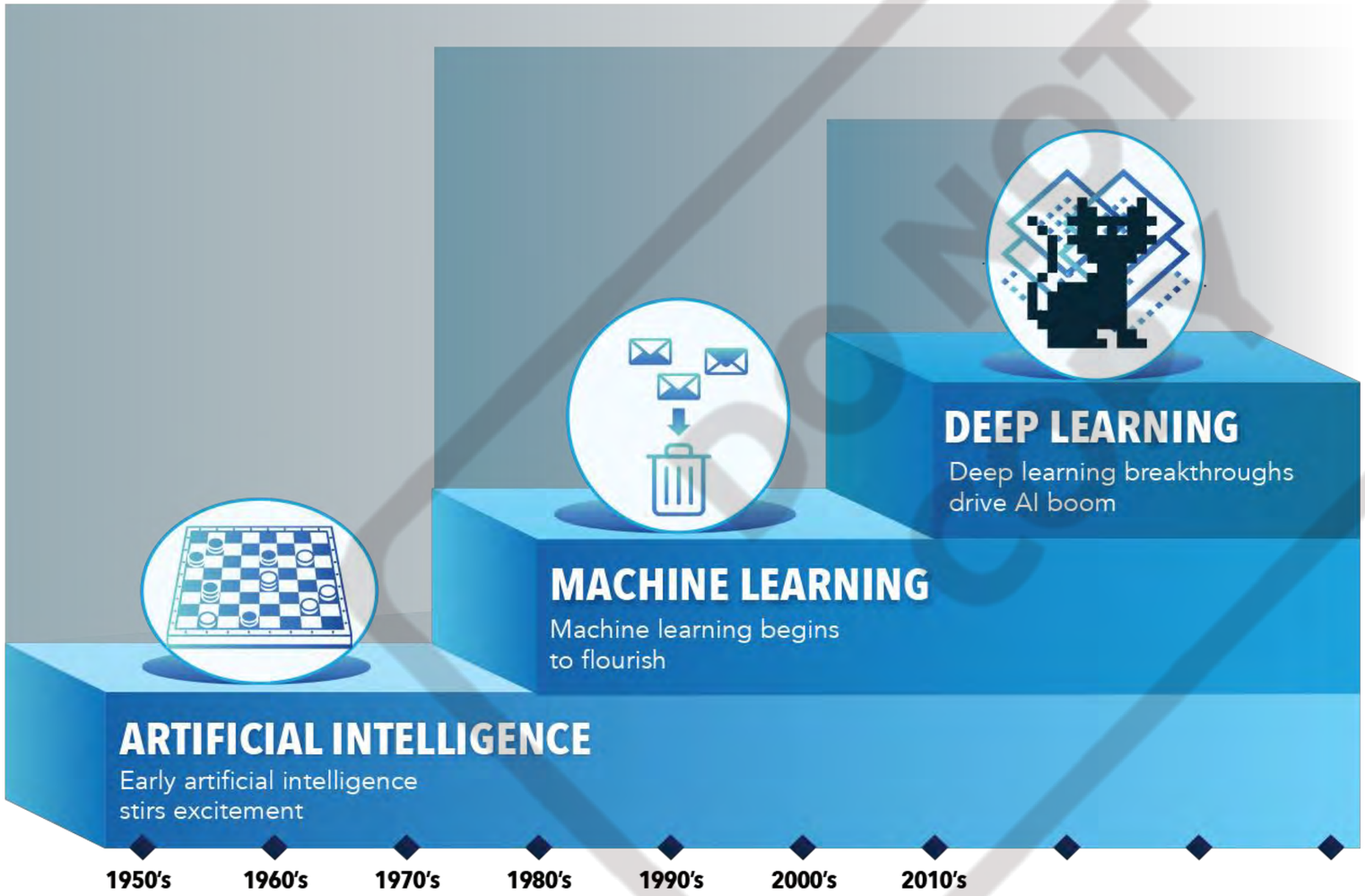


Poses serious analytical challenges

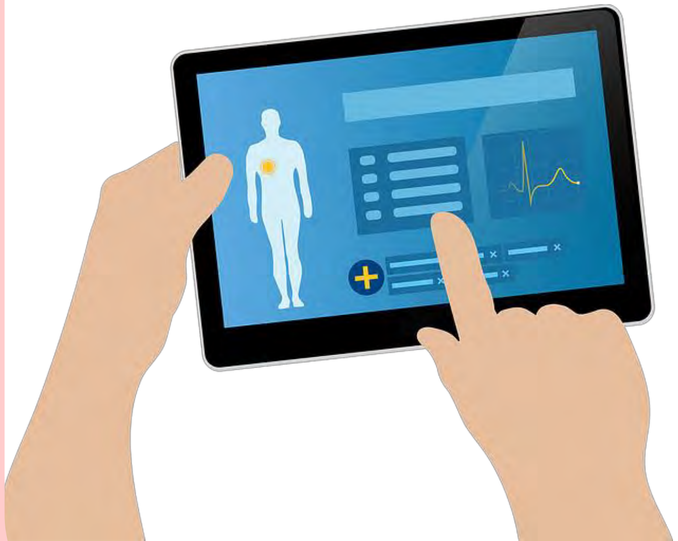
- ❑ High dimensionality
- ❑ Need to mine complex patterns of interactions
- ❑ Multicollinearity
- ❑ Heterogenous data
- ❑ Low sample size

Machine learning & AI tools

AI &
CLINICAL DECISION SUPPORT SYSTEMS

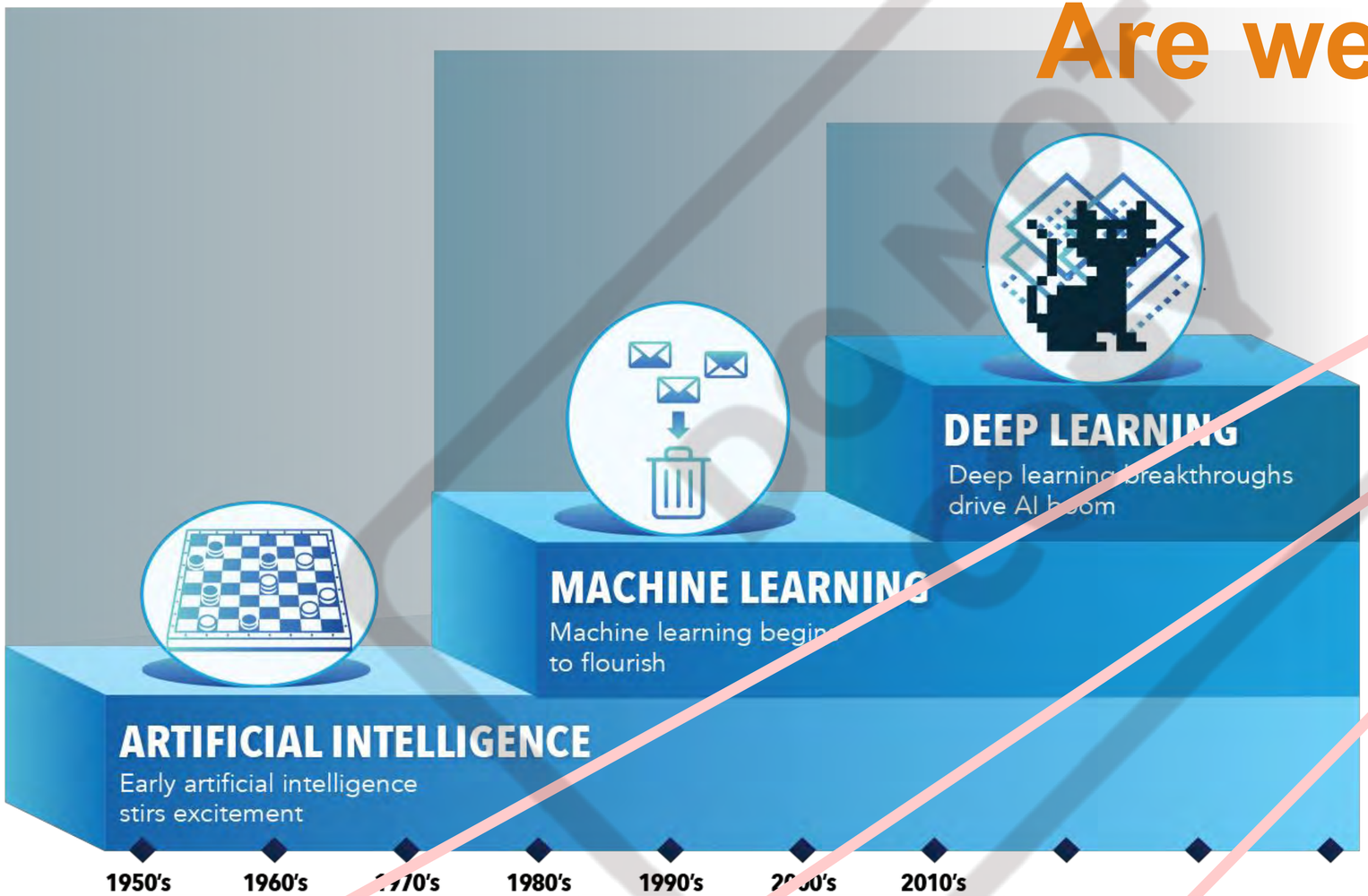


A CLINICAL DECISION SUPPORT SYSTEM is a computerized program that supports determinations, judgments, and courses of action in healthcare

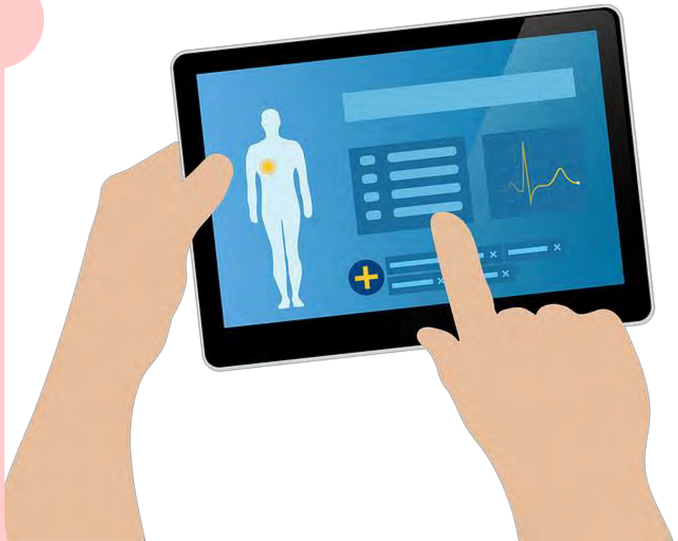


AI &
CLINICAL DECISION SUPPORT SYSTEMS

Are we ready?



A CLINICAL DECISION SUPPORT SYSTEM is a computerized program that supports determinations, judgments, and courses of action in healthcare

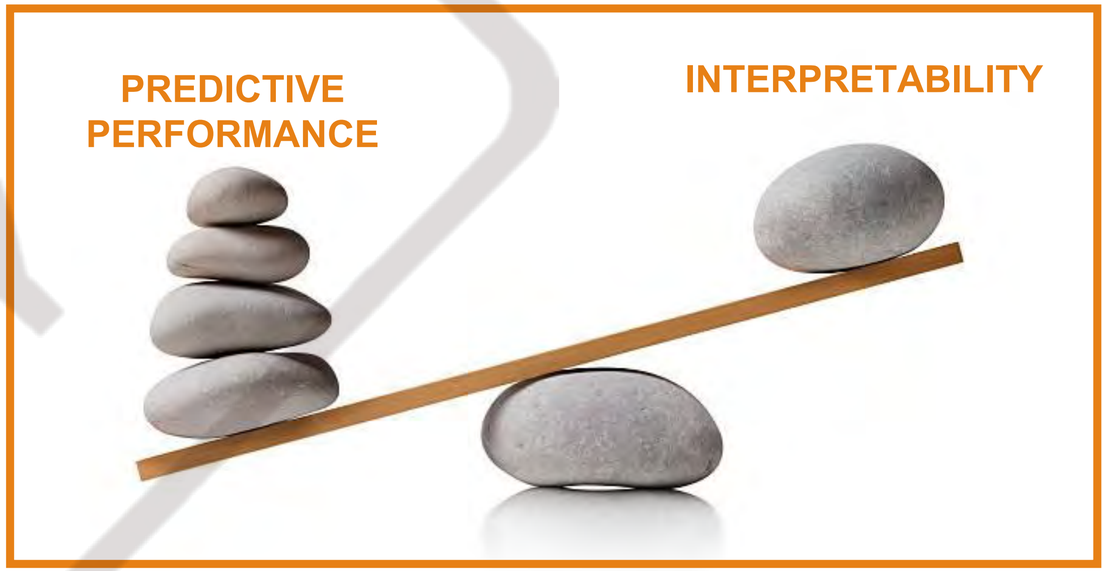


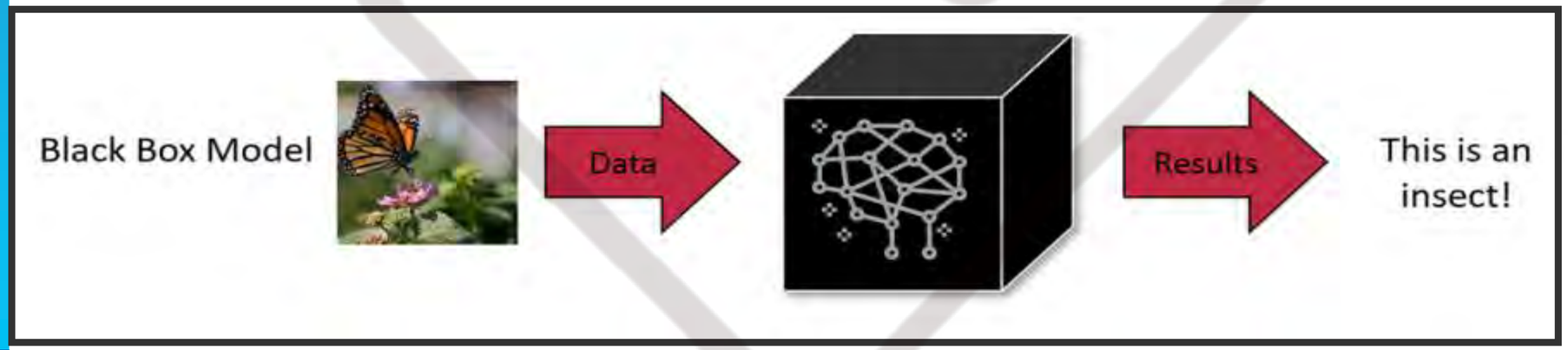
Trustworthy | **Fair & Ethics** | **eXplicable**

THE BLACK BOX PARADOX: prediction vs eXplainability

Many ML techniques can integrate high-dim data from different information sources, with high predictive performances.

Nevertheless, often, prediction is achieved at the cost of their interpretability

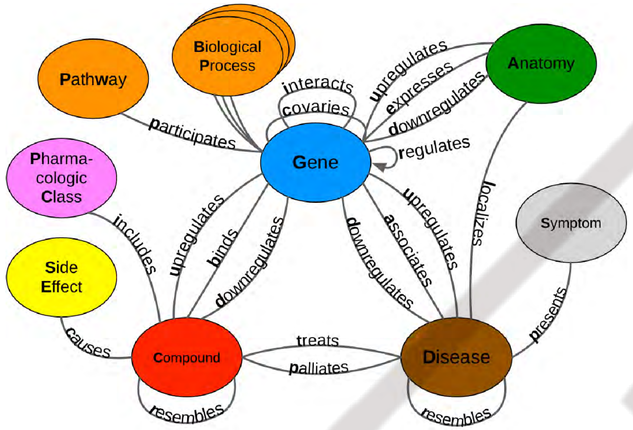




We see the inputs and the predictions (output), but the processes in between remain hidden

Machine Learning: prediction vs eXplainability

In healthcare, otherwise, when we build predictive models, we are interested not only in predicting well the outcome but also in understanding the model, which can help us generate new knowledge:



- **Importance of predictors**
- **Directionality of associations**

Especially necessary in the case of multi-modal data

The XAI trend…

recommends using ML models whose nature is self-explanatory

Contents lists available at ScienceDirect

Information Fusion

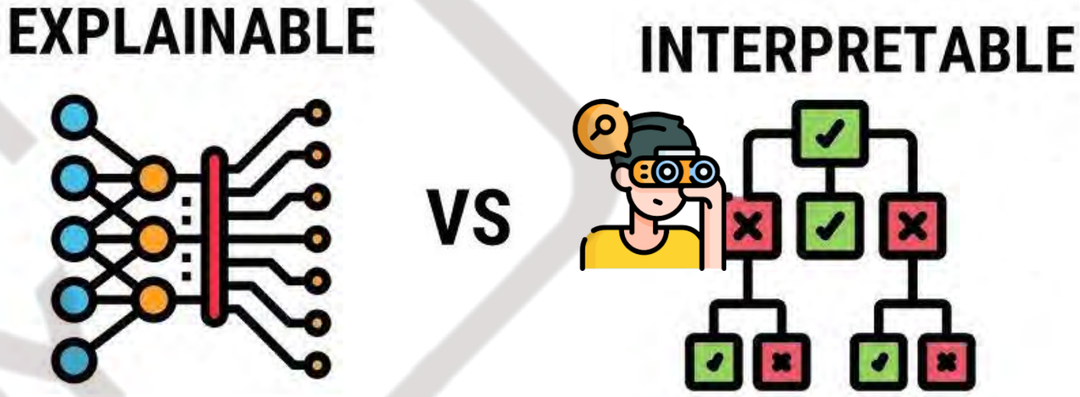
journal homepage: www.elsevier.com/locate/inffus

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

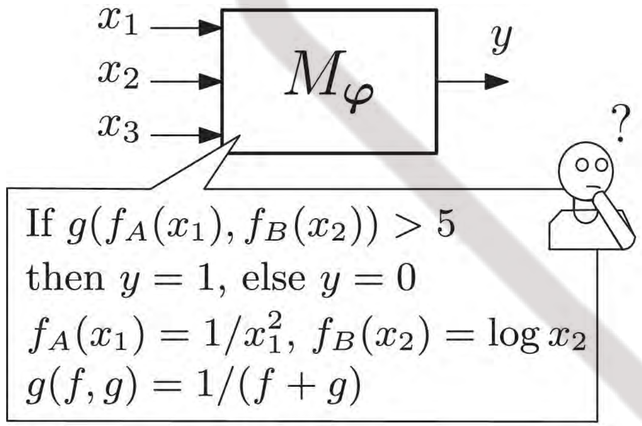
Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[a,c,d,*], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[h], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], Francisco Herrera[g]

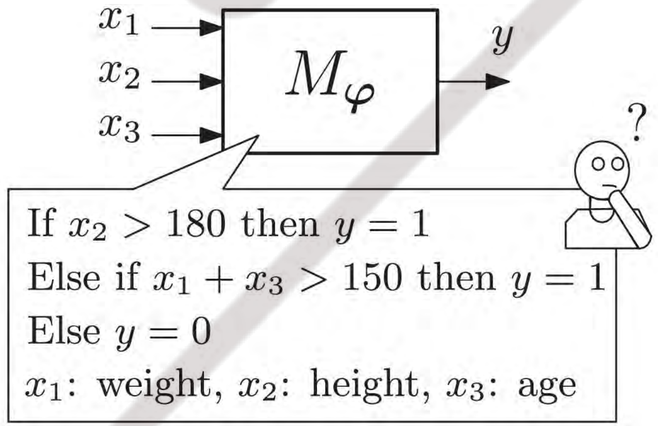
XAI: concepts, taxonomies & opportunities

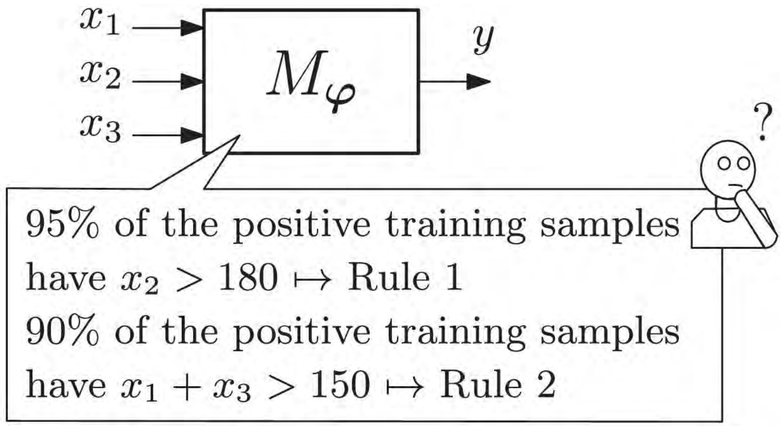
- **eXplainability** refers to the degree at which a given model tries to clarify or detail its internal functions
- **Interpretability** refers to the level at which a given model makes sense for a human Observer (also known as transparency)



THREE EXPLAINABLE MODELS WITH DIFFERENT LEVELS OF INTERPRETABILITY/TRANSPARENCY







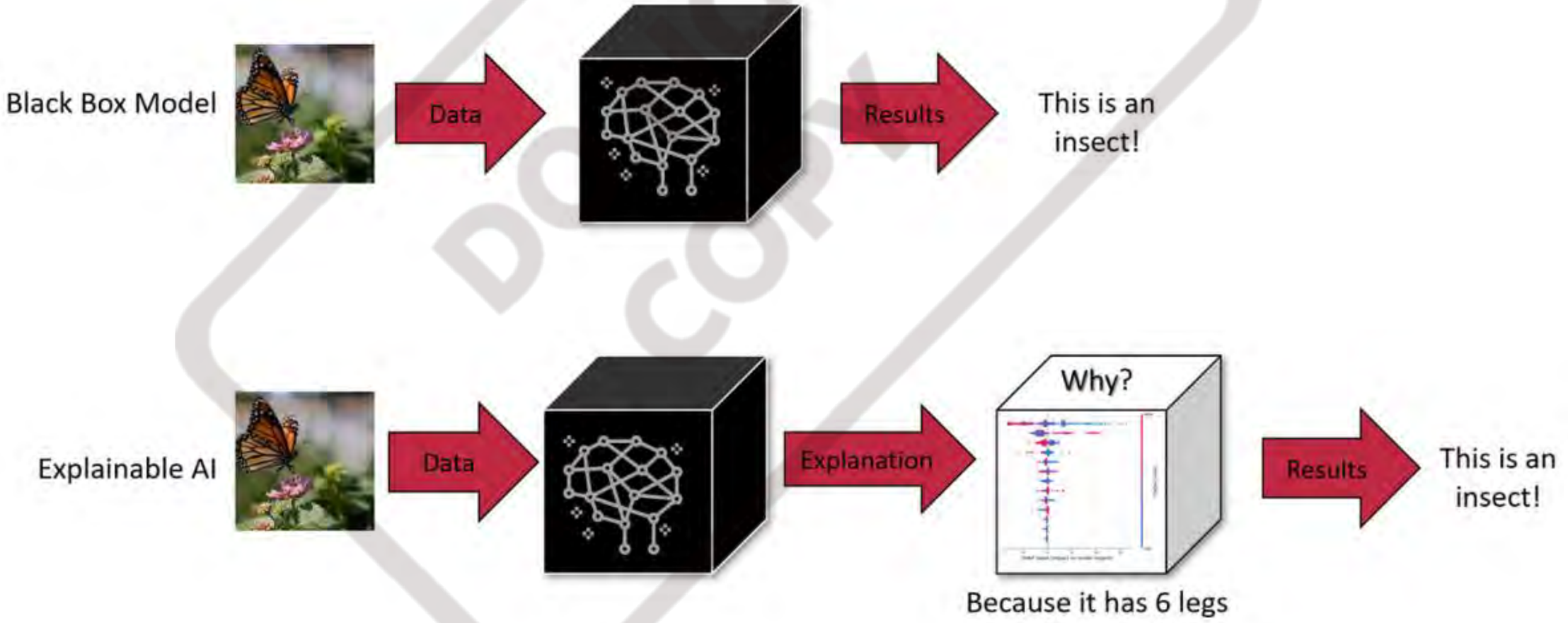
XAI for healthcare

Interpretability is A MUST if talk about CLINICAL APPLICATIONS

In many cases, it is more important to understand "how the decision was made" than the decision itself…

XAI: POST-HOC EXPLAINERS

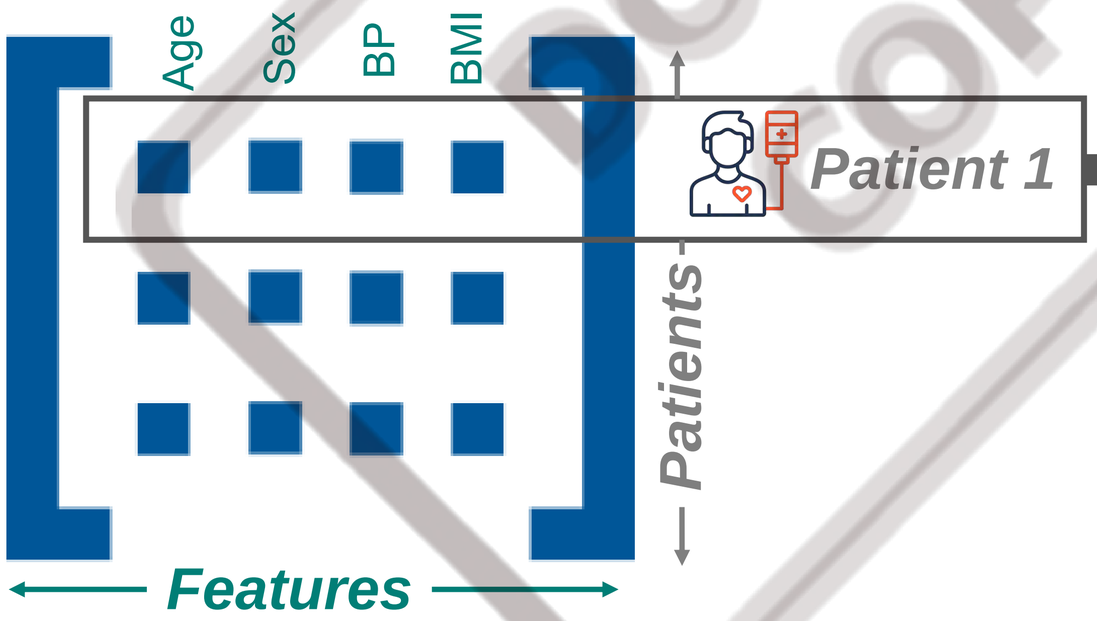
SOMETIMES, MODELS ARE NOT INTERPRETABLE/TRANSPARENT BY THEMSELVES BUT WE CAN MAKE THEM SO USING POST-HOC EXPLAINERS



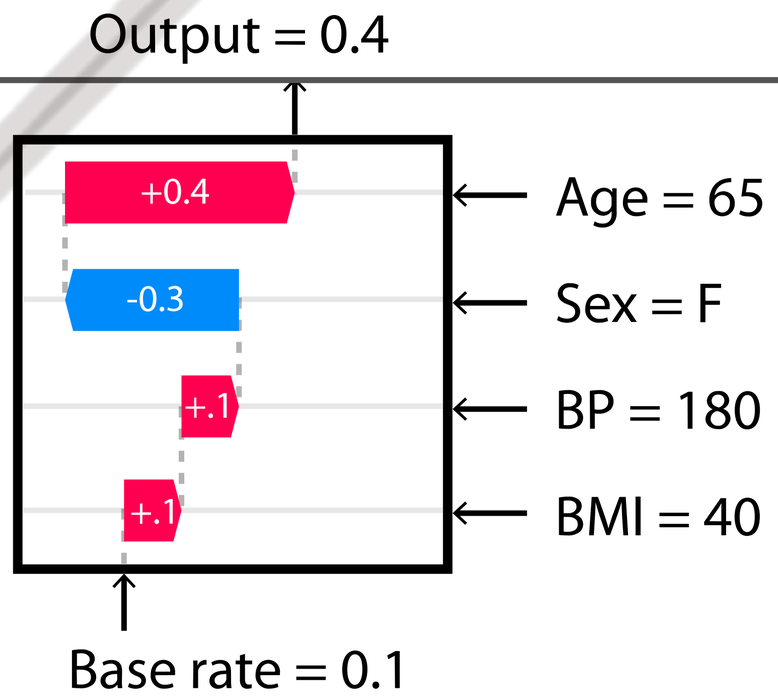
SHAP values (SHapley Additive exPlanations)

SHAP values (SHapley Additive exPlanations) is a method based on cooperative game theory and used to increase transparency and interpretability of black box algorithms

SHAP produces a matrix with the individual contribution of each feature for each example or observation



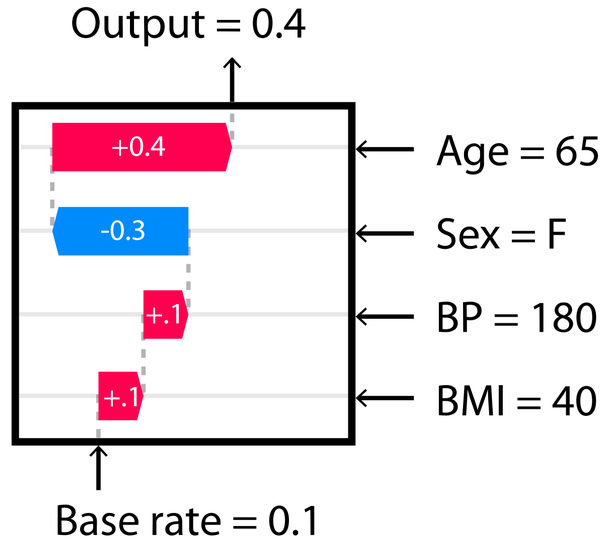
Cardiovascular risk prediction for a specific patient will result from the sum of all feature contributions



SHAP values (SHapley Additive exPlanations)

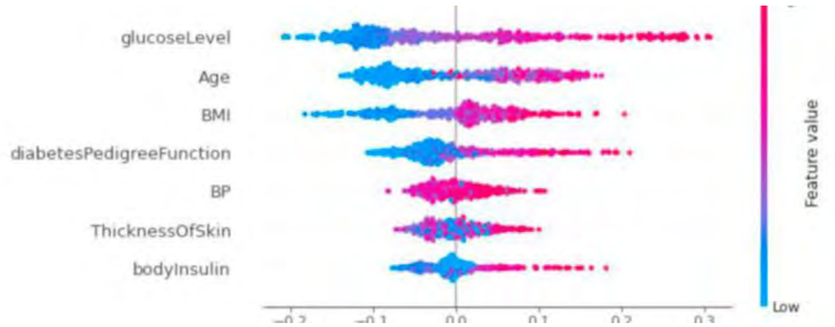
LOCAL LEVEL EXPLANATIONS:

- Understanding contributing **risk factor for specific individuals** or population subgroups
- Identifying the **reason** why the model **failed prediction** in some individuals

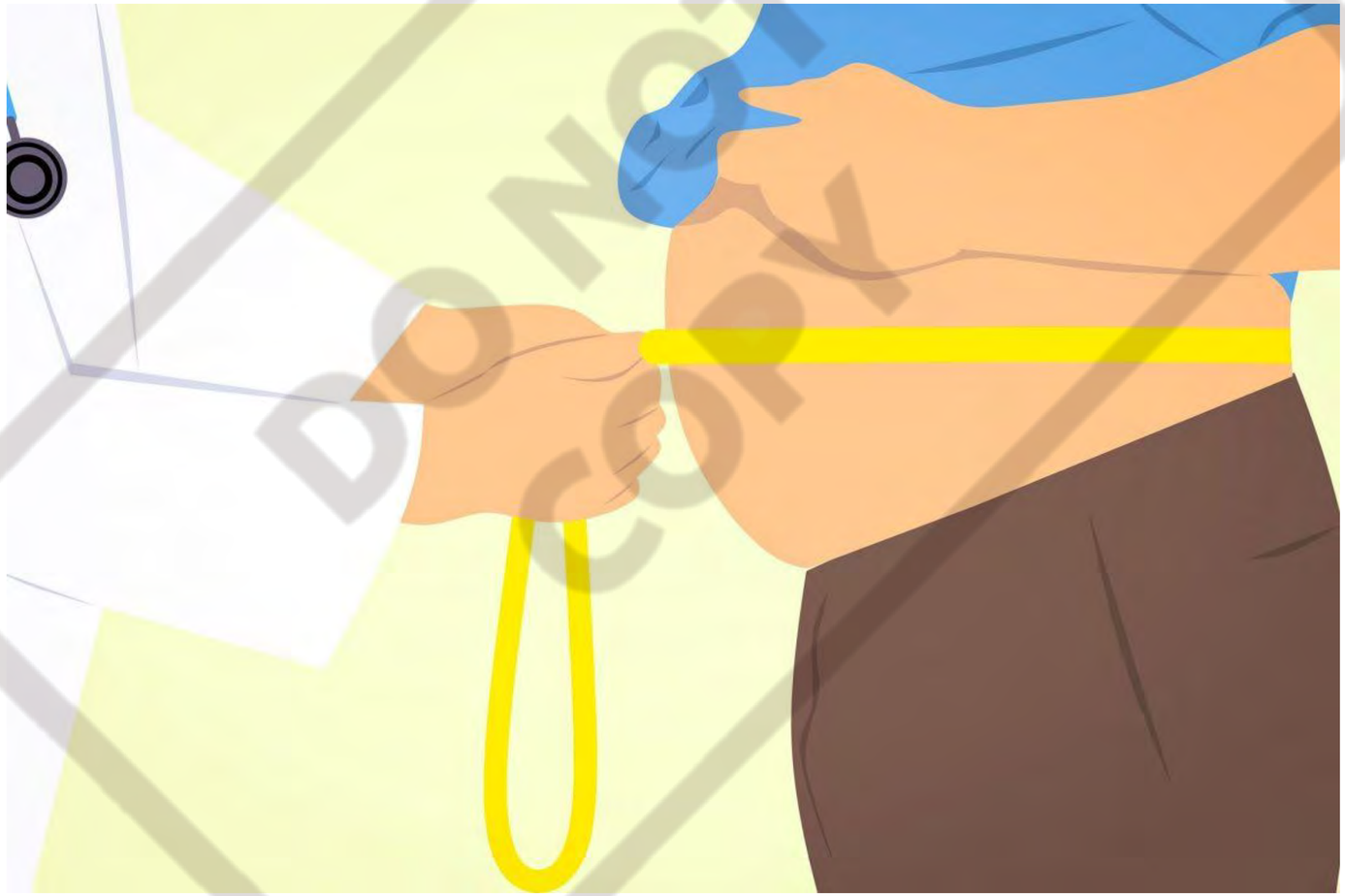


GLOBAL LEVEL EXPLANATIONS:

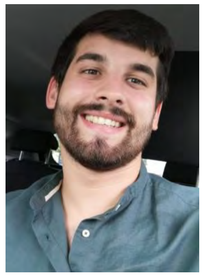
- **Feature importance** (ranking of variables)
- **Directionality** of associations



CASE STUDY: A real story on how to use XAI for helping children with obesity

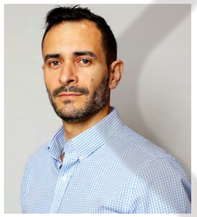


CASE STUDY:



MSc Álvaro Torres Martos

Dra. Aguilera-García Concepción



Dr. Jesús Alcalá-Fdez

All codes available in

GitHub

https://github.com/AlvaroTorresMartos/IR_prediction

Artificial Intelligence In Medicine 156 (2024) 102962

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study

Álvaro Torres-Martos [a,b,c,1], Augusto Anguita-Ruiz [c,d,1], Mireia Bustos-Aibar [a,c,e], Alberto Ramírez-Mena [f], María Arteaga [g], Gloria Bueno [c,e,h], Rosaura Leis [c,i], Concepción M. Aguilera [a,b,c,*], Rafael Alcalá [g], Jesús Alcalá-Fdez [g]

[a] Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain
[b] Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, 18012, Spain
[c] CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, 28029, Spain
[d] Barcelona Institute for Global Health, ISGlobal, Barcelona, 08003, Spain
[e] Growth, Exercise, Nutrition and Development (GENUD) Research Group, Institute for Health Research Aragón (IIS Aragón), Zaragoza, 50009, Spain
[f] Bioinformatics Unit, Centre for Genomics and Oncological Research, GENYO (Pfizer/University of Granada/Andalusian Regional Government, PTS, Granada, 18016, Spain
[g] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain
[h] Pediatric Endocrinology Unit, Facultad de Medicina, Clínic University Hospital Lozano Blesa, University of Zaragoza, Zaragoza, 50009, Spain
[i] Unit of Pediatric Gastroenterology, Hepatology and Nutrition, Pediatric Service, Hospital Clínico Universitario de Santiago. Unit of Investigation in Nutrition, Growth and Human Development of Galicia-USC, Pediatric Nutrition Research Group-Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, 15706, Spain

ARTICLE INFO

Keywords:
Pediatric obesity
Insulin resistance
Epigenomics
Multiomics
Machine Learning
Explainable Artificial Intelligence

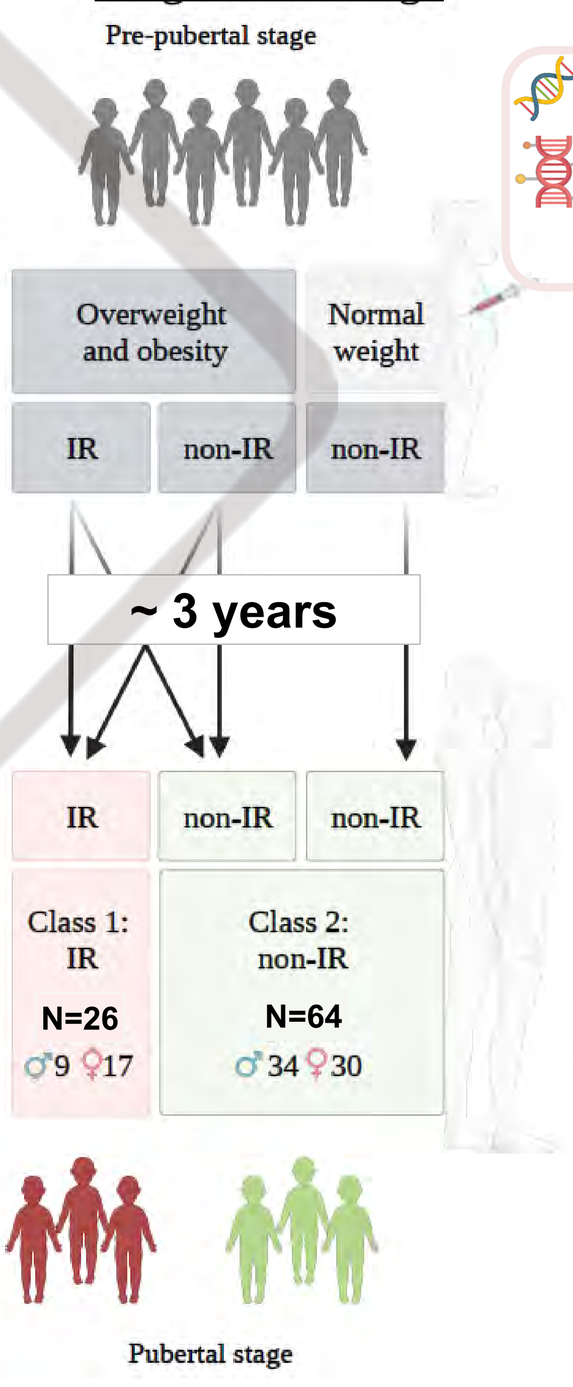
ABSTRACT

Pediatric obesity can drastically heighten the risk of cardiometabolic alterations later in life, with insulin resistance standing as the cornerstone linking adiposity to the increased cardiovascular risk. Puberty has been pointed out as a critical stage after which obesity associated insulin resistance is more difficult to revert. Timely prediction of insulin resistance in pediatric obesity is therefore vital for mitigating the risk of its associated comorbidities. The construction of effective and robust predictive systems for a complex health outcome like insulin resistance during the early stages of life demands the adoption of longitudinal designs for more causal inferences, and the integration of factors of varying nature involved in its onset. In this work, we propose an eXplainable Artificial Intelligence-based decision support pipeline for early diagnosis of insulin resistance in a longitudinal cohort of 90 children. For that, we leverage multi-omics (genomics and epigenomics) and clinical data from the pre-pubertal stage. Different data layers combinations, pre-processing techniques (missing values, feature selection, class imbalance, etc.), algorithms, training procedures were considered following good practices for Machine Learning. SHapley Additive exPlanations were provided for specialists to understand both the decision-making mechanisms of the system and the impact of the features on each automatic decision, an essential issue in high-risk areas such as this one where system decisions may affect people's lives. The system showed a relevant predictive ability (AUC and G-mean of 0.92). A deep exploration, both at the global and the local level, revealed promising biomarkers of insulin resistance in our population, highlighting classical markers, such as Body Mass Index z-score or leptin/adiponectin ratio, and novel ones such as methylation patterns of relevant genes, such as *HDAC4*, *PTPRN2*, *MATN2*, *RASGRF1* and *EBF1*. Our findings highlight the importance of integrating multi-omics data and following eXplainable Artificial Intelligence trends when building decision support systems.

* Corresponding author at: Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain.
E-mail addresses: alvarotorres@ugr.es (Á. Torres-Martos), augusto.anguita@isglobal.org (A. Anguita-Ruiz), mbustos@iisaragon.es (M. Bustos-Aibar), alberto.ramirez@genyo.es (A. Ramírez-Mena), mariaartj@correo.ugr.es (M. Arteaga), mgbuenol@unizar.es (G. Bueno), mariarosaura.leis@usc.es (R. Leis), caguiler@ugr.es (C.M. Aguilera), alcala@decsai.ugr.es (R. Alcalá), jalcala@decsai.ugr.es (J. Alcalá-Fdez).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.artmed.2024.102962
Received 27 February 2024; Received in revised form 31 July 2024; Accepted 16 August 2024
Available online 20 August 2024

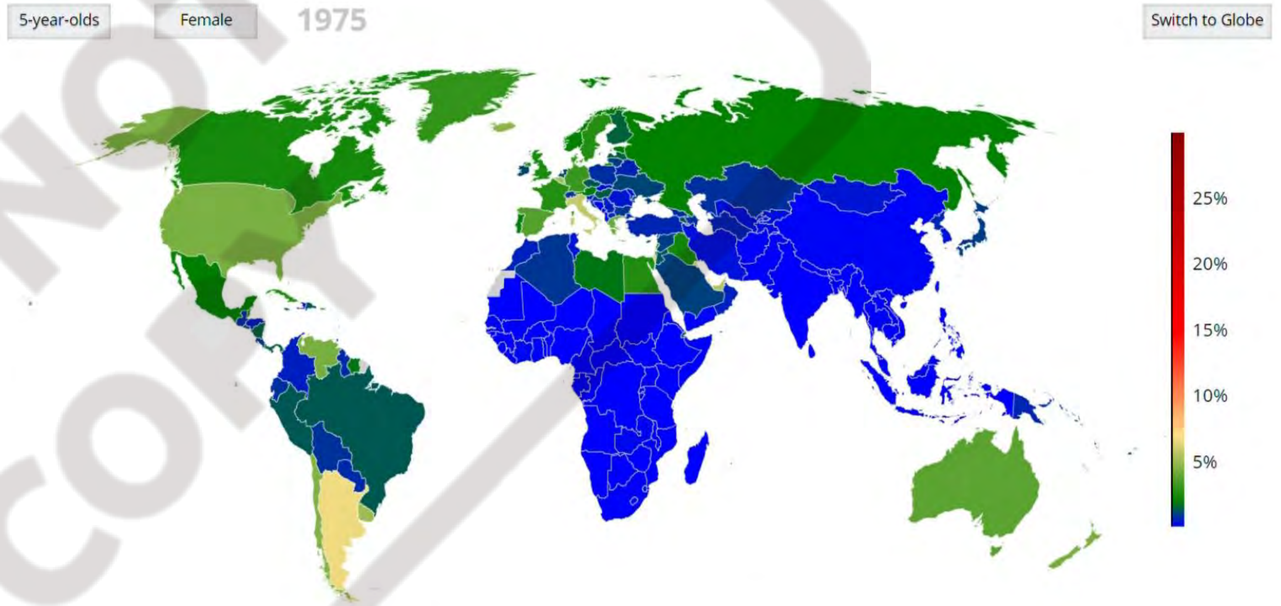
Longitudinal design



CASE STUDY: Research context

Some WHO Key facts…

- The obesity problem has grown to **pandemic proportions**
- The **prevalence** of **overweight and obesity** among children aged 5-19 has **risen** dramatically from just 4% in 1975 to just over **20% in 2022**

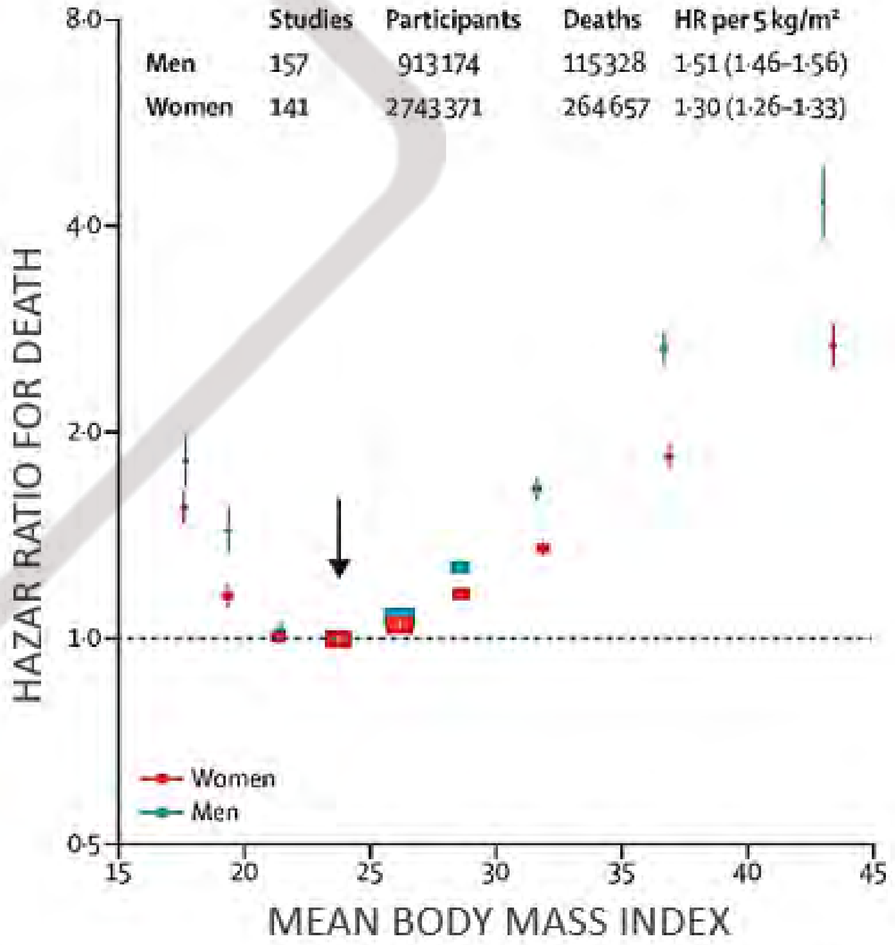


CASE STUDY: Research context

Raised childhood BMI is a major **risk factor** for **noncommunicable diseases during adulthood** such as...

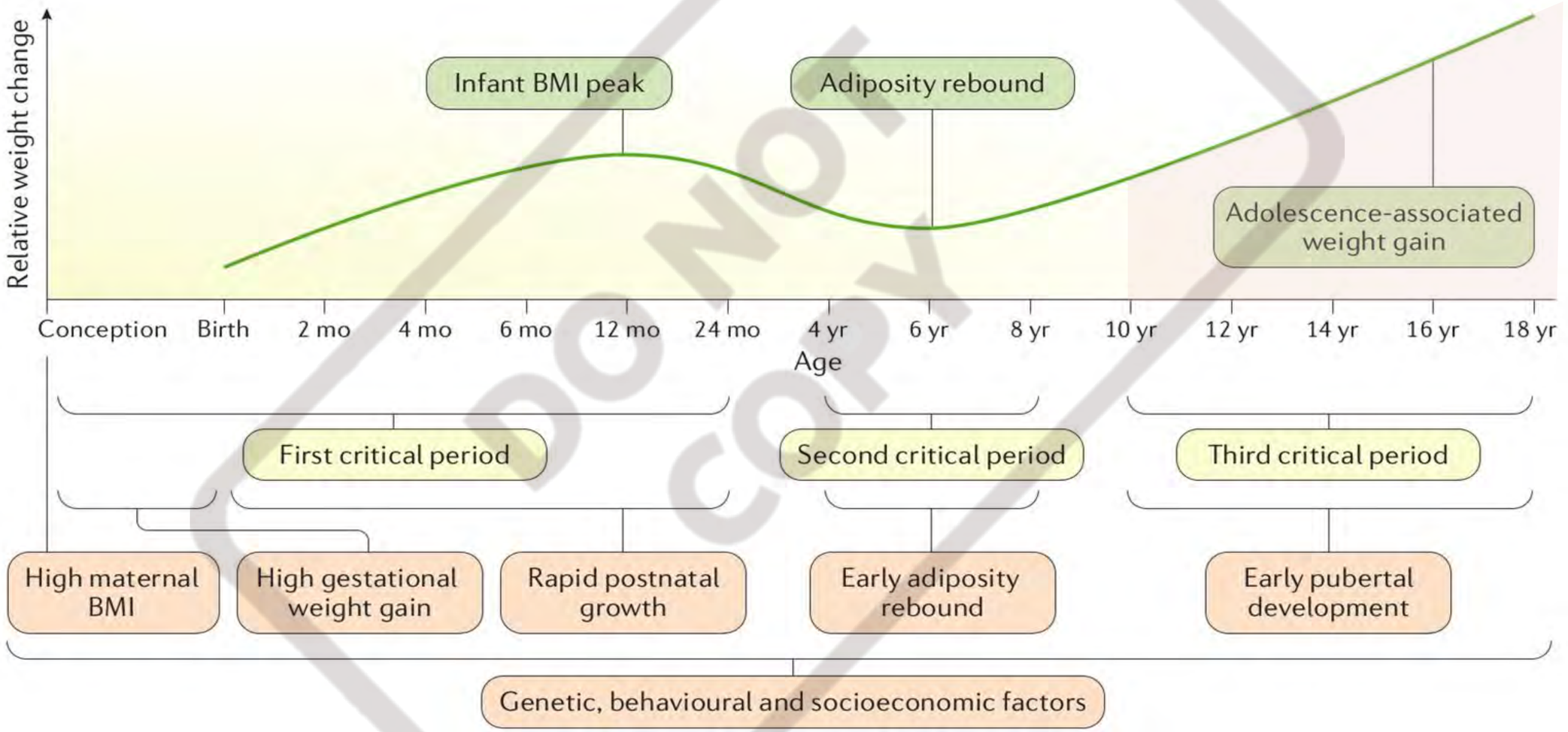
- Cardiovascular diseases
- Diabetes
- Musculoskeletal disorders
- Cancers

The **risk** of these and other **noncommunicable diseases increases** even when a person is only slightly overweight and grows more serious as the **BMI rise**..



Data sources: Global BMI Mortality Collaboration. **Lancet**. 2016;388(10046):776-786. doi:10.1016/S0140-6736(16)30175-1

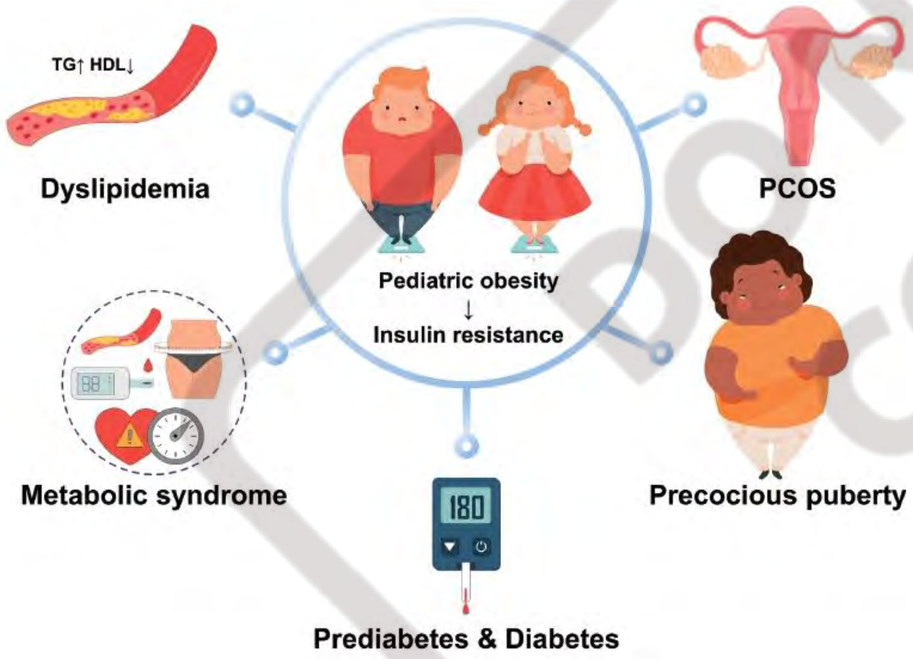
CASE STUDY: Research context



González-Muniesa, P, et al. Obesity. Nature Reviews Disease Primers. 2017;15;3:17034. doi: 10.1038/nrdp.2017.34.

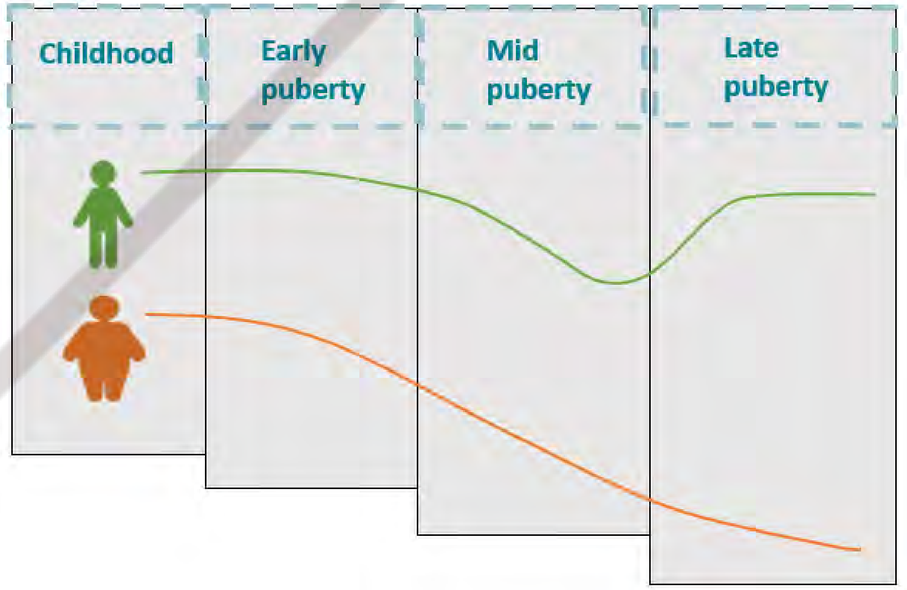
CASE STUDY: Research context

Pediatric obesity can drastically heighten the risk of cardiometabolic alterations later in life, with insulin resistance (IR) standing as the cornerstone linking adiposity to the increased cardiovascular risk

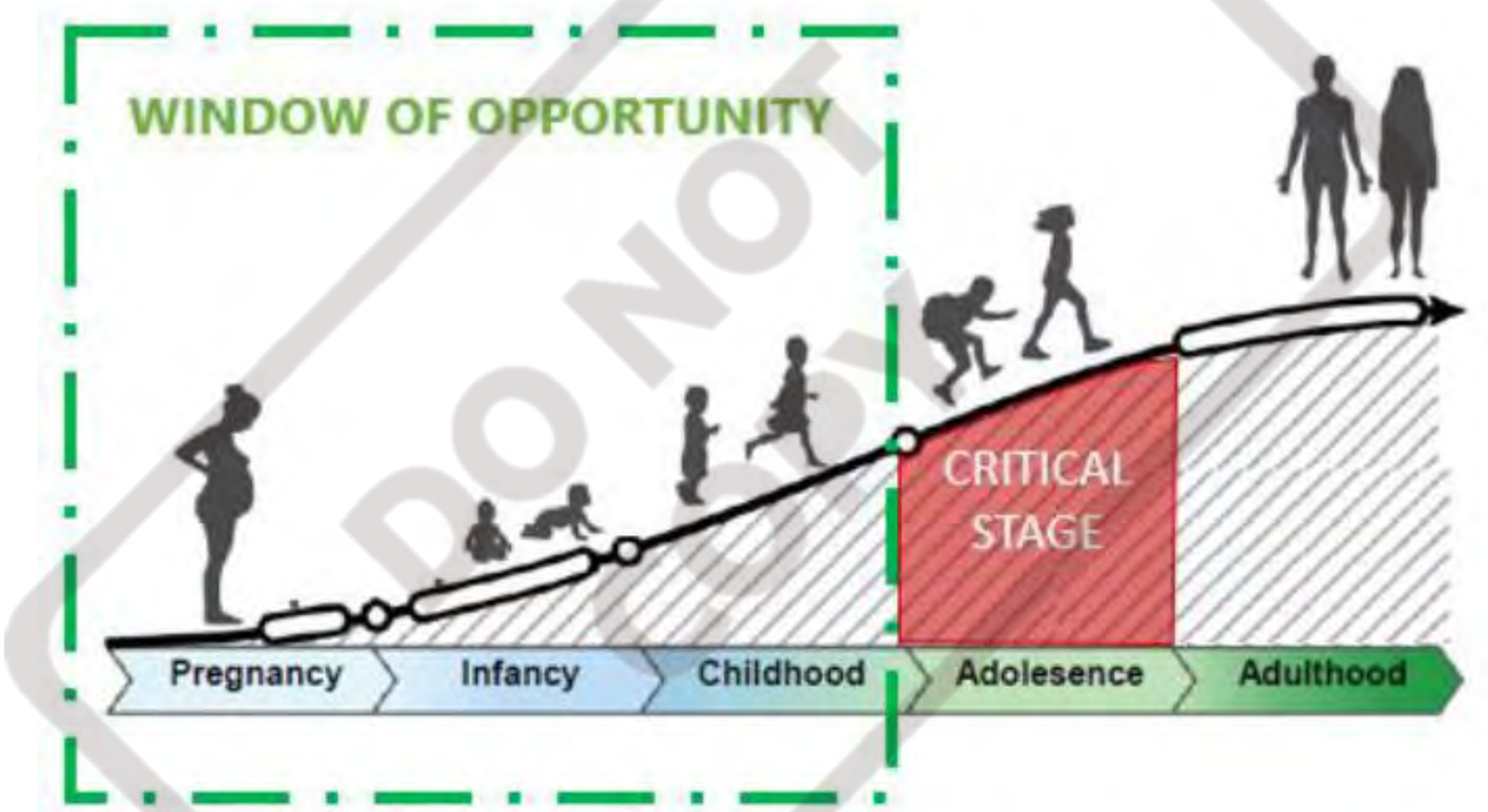


Puberty has been pointed out as a critical stage upon which obesity-associated IR is more difficult to revert

Insulin sensitivity trajectories in
normal weight *vs.* **children with obesity**



CASE STUDY: Research context



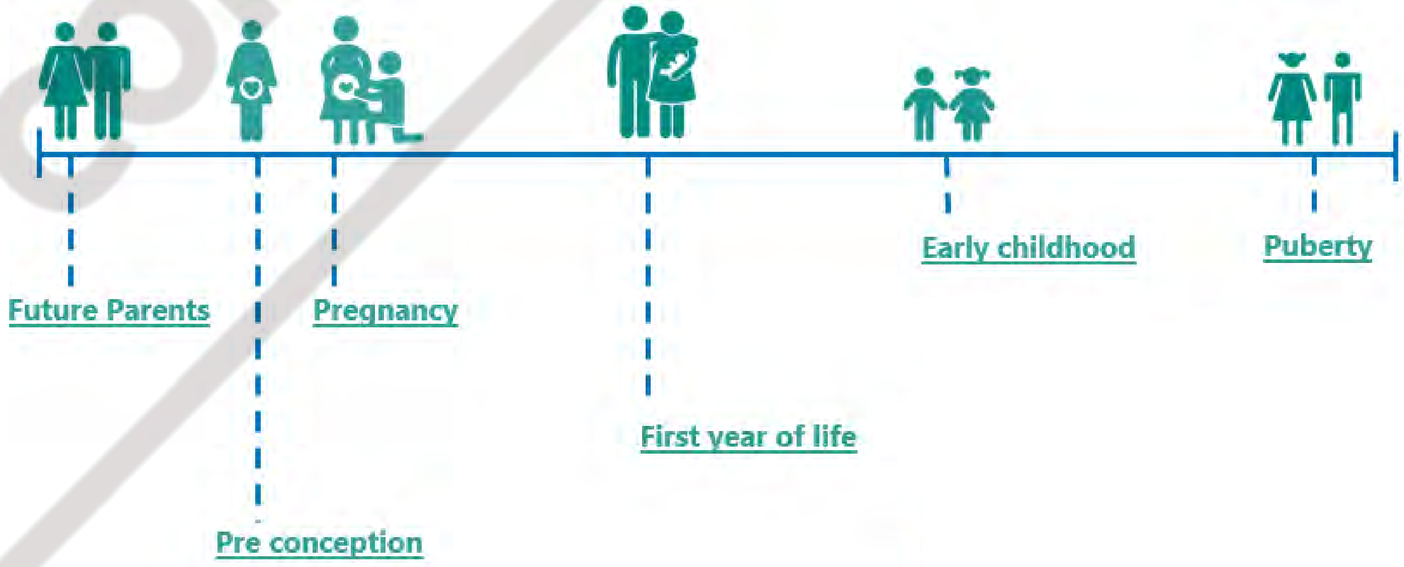
Early childhood appears as a magnificent window of opportunity for the implementation of preventive actions against obesity-associated IR worsening and appearance

CASE STUDY: insulin resistance prediction is not a trivial task

The integration of factors of varying nature involved in its onset (i.e., genetics, epigenetics, proteins, clinical and endogenous factors, and of course the environment)

Normally, these layers are high-dimensional

The adoption of longitudinal designs



Artificial Intelligence In Medicine 156 (2024) 102962

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study

Álvaro Torres-Martos [a,b,c,1], Augusto Anguita-Ruiz [c,d,1], Mireia Bustos-Aibar [a,c,e], Alberto Ramírez-Mena [f], María Arteaga [g], Gloria Bueno [c,e,h], Rosaura Leis [c,i], Concepción M. Aguilera [a,b,c,*], Rafael Alcalá [g], Jesús Alcalá-Fdez [g]

[a] Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain
[b] Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, 18012, Spain
[c] CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, 28029, Spain
[d] Barcelona Institute for Global Health, ISGlobal, Barcelona, 08003, Spain
[e] Growth, Exercise, Nutrition and Development (GENUD) Research Group, Institute for Health Research Aragón (IIS Aragón), Zaragoza, 50009, Spain
[f] Bioinformatics Unit, Centre for Genomics and Oncological Research, GENYO Pfizer/University of Granada/Andalusian Regional Government, PTS, Granada, 18016, Spain
[g] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain
[h] Pediatric Endocrinology Unit, Facultad de Medicina, Clinic University Hospital Lozano Blesa, University of Zaragoza, Zaragoza, 50009, Spain
[i] Unit of Pediatric Gastroenterology, Hepatology and Nutrition, Pediatric Service, Hospital Clínico Universitario de Santiago. Unit of Investigation in Nutrition, Growth and Human Development of Galicia-USC, Pediatric Nutrition Research Group-Health Research Institute of Santiago de Compostela (IDIS), Santiago de Compostela, 15706, Spain

ARTICLE INFO

Keywords:
Pediatric obesity
Insulin resistance
Epigenomics
Multiomics
Machine Learning
Explainable Artificial Intelligence

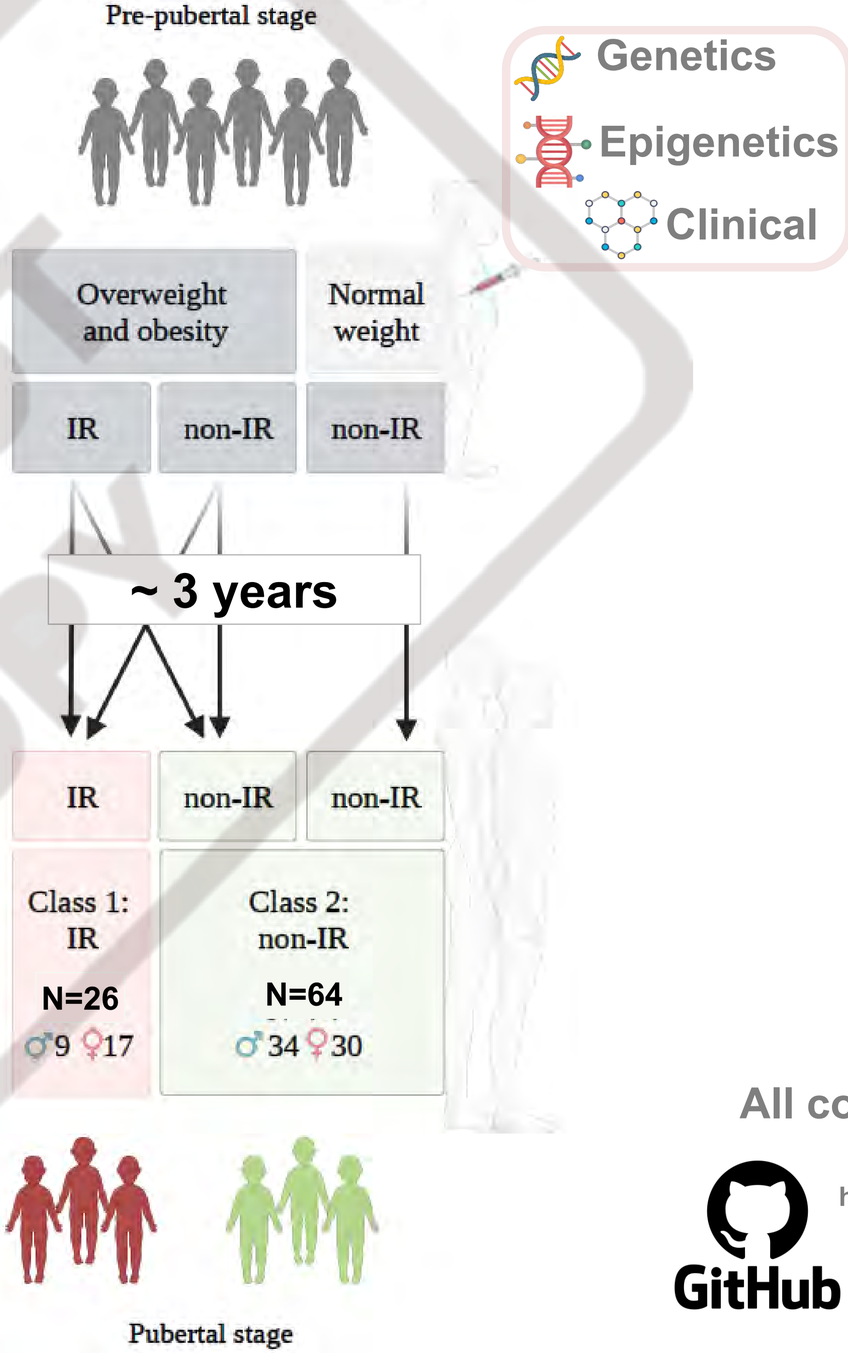
ABSTRACT

Pediatric obesity can drastically heighten the risk of cardiometabolic alterations later in life, with insulin resistance standing as the cornerstone linking adiposity to the increased cardiovascular risk. Puberty has been pointed out as a critical stage after which obesity-associated insulin resistance is more difficult to revert. Timely prediction of insulin resistance in pediatric obesity is therefore vital for mitigating the risk of its associated comorbidities. The construction of effective and robust predictive systems for a complex health outcome like insulin resistance during the early stages of life demands the adoption of longitudinal designs for more causal inferences, and the integration of factors of varying nature involved in its onset. In this work, we propose an eXplainable Artificial Intelligence-based decision support pipeline for early diagnosis of insulin resistance in a longitudinal cohort of 90 children. For that, we leverage multi-omics (genomics and epigenomics) and clinical data from the pre-pubertal stage. Different data layers combinations, pre-processing techniques (missing values, feature selection, class imbalance, etc.), algorithms, training procedures were considered following good practices for Machine Learning. SHapley Additive exPlanations were provided for specialists to understand both the decision-making mechanisms of the system and the impact of the features on each automatic decision, an essential issue in high-risk areas such as this one where system decisions may affect people's lives. The system showed a relevant predictive ability (AUC and G-mean of 0.92). A deep exploration, both at the global and the local level, revealed promising biomarkers of insulin resistance in our population, highlighting classical markers, such as Body Mass Index z-score or leptin/adiponectin ratio, and novel ones such as methylation patterns of relevant genes, such as *HDAC4*, *PTPRN2*, *MATN2*, *RASGRF1* and *EBF1*. Our findings highlight the importance of integrating multi-omics data and following eXplainable Artificial Intelligence trends when building decision support systems.

* Corresponding author at: Department of Biochemistry and Molecular Biology II, School of Pharmacy, "José Mataix Verdú" Institute of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, University of Granada, Granada, 18071, Spain.
E-mail addresses: alvarotorres@ugr.es (Á. Torres-Martos), augusto.anguita@isglobal.org (A. Anguita-Ruiz), mbustos@iisaragon.es (M. Bustos-Aibar), alberto.ramirez@genyo.es (A. Ramírez-Mena), mariaarj@correo.ugr.es (M. Arteaga), mgbueno@unizar.es (G. Bueno), mariarosaura.leis@usc.es (R. Leis), caguiler@ugr.es (C.M. Aguilera), alcala@decsai.ugr.es (R. Alcalá), jalcala@decsai.ugr.es (J. Alcalá-Fdez).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.artmed.2024.102962
Received 27 February 2024; Received in revised form 31 July 2024; Accepted 16 August 2024
Available online 20 August 2024

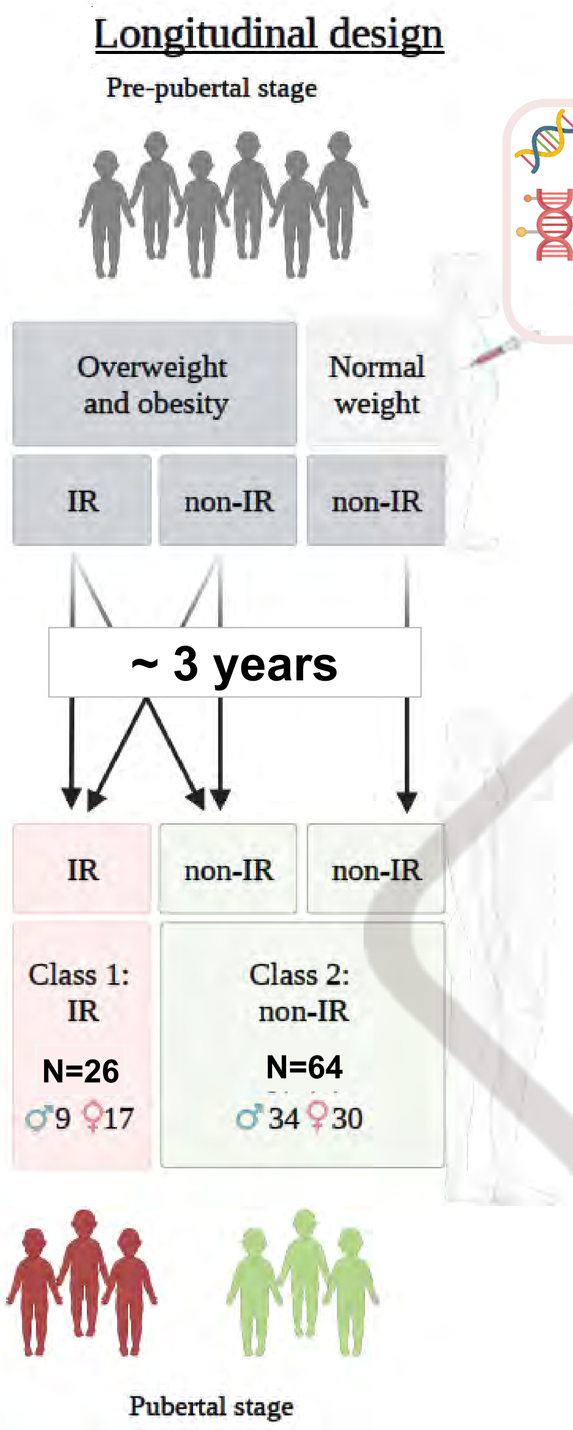
Longitudinal design

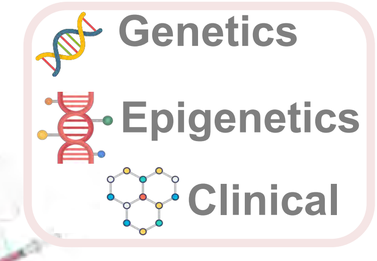


All codes available in

GitHub

https://github.com/AlvaroTorresMartos/IR_prediction





GOOD PRACTICES FOR GENERATING CDSSs WITH MULTI-OMICS & CLINICAL DATA

DATA PRE-PROCESSING

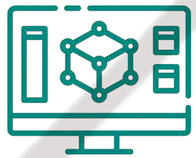
- Outliers treatment

- Missing data

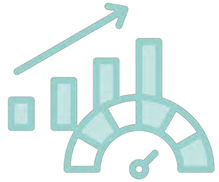
- Data dimensionality

ML MODEL CONSTRUCTION & DATA FUSION

- Model selection



- Performance metrics

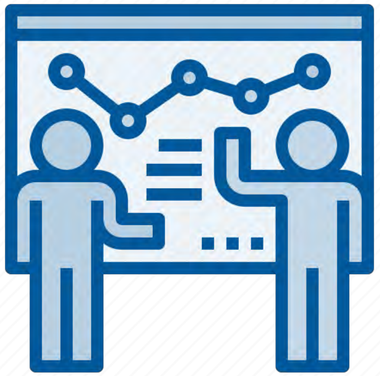


- Data fusion

- Validation

MODEL INTERPRETATION

- How to use POST-HOC explainers



Final Model explanation

A final model was generated using the whole dataset for interpretation

RandomForest

Predictive ability of the final classifier

	Classification metrics				
Method	Accuracy	Sensitivity	Specificity	AUC	G-mean
RF	0.90	1.00	0.85	0.92	0.92

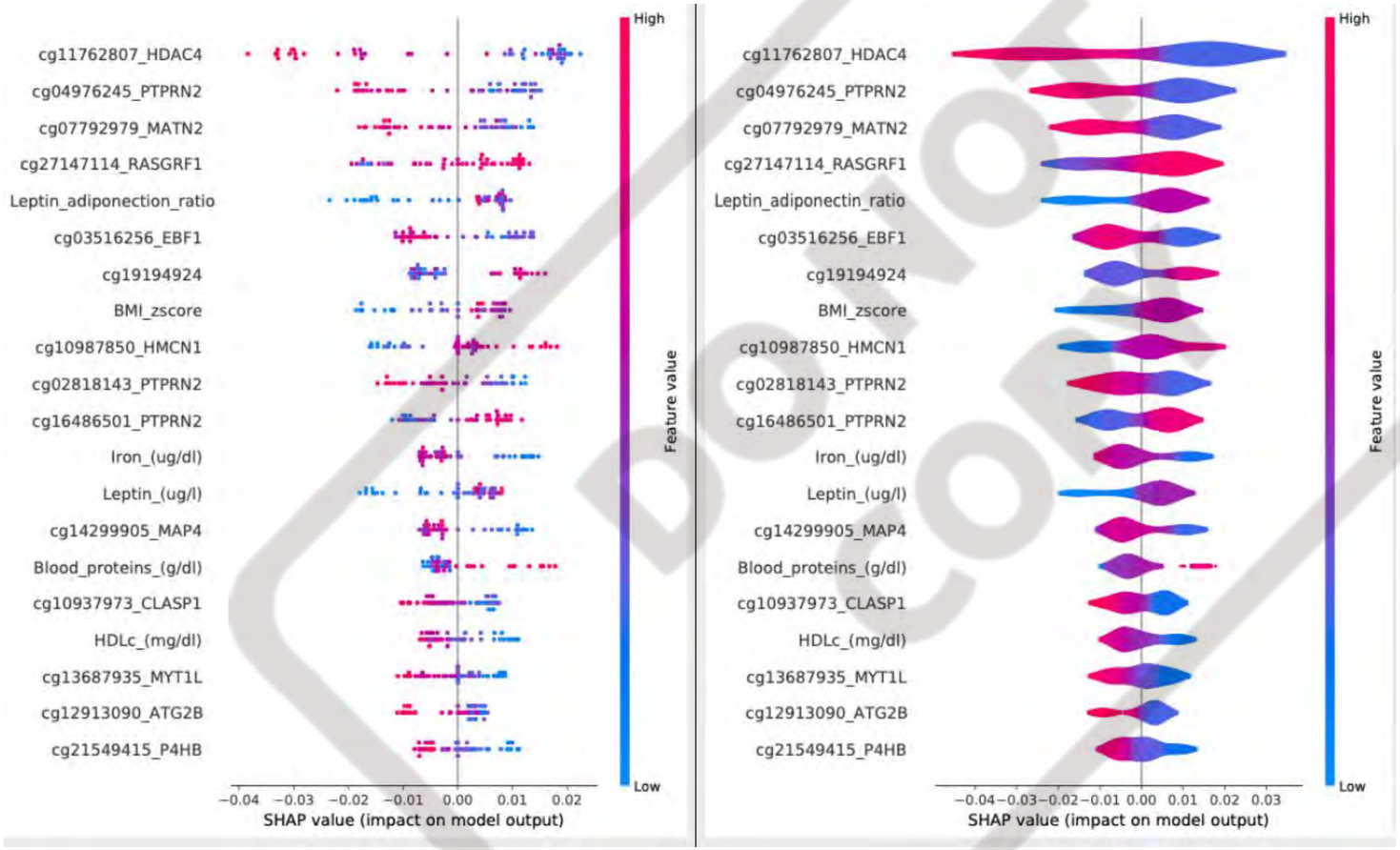
Breiman rated RF as A+ for prediction but F for interpretability

SHAP

A sensitivity value of 1 indicates that the system has correctly predicted all true positives, without false negatives

Final Model explanation: Knowledge extraction

Global explanations

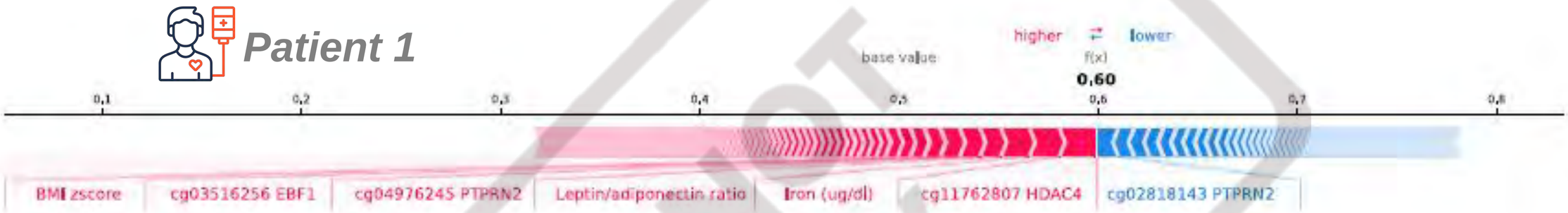


- Each individual/patient is a dot
- Combination of biomarkers from both layers in the top ranking
- Directionality of associations

Although RF uses certain variables more frequently than others to make predictions, it is important to note that IT IS THE SUM ALL feature SHAP values what determines the prediction towards one class or another in each patient

Final Model explanation: Personalized intervention

- We can go deeper into the explanations (at the level of individuals or small groups of them)



- If we identify which are the risk factors for a specific individual, we might define personalized intervention plans

1. Start intervention to induce hypermethylation of HDAC4 (if any)
2. Treating iron deficiency
3. Anti-inflammatory intervention to revert leptin dysregulation

Final Model explanation: Clinical thresholds

Local explanations

The study of SHAP values for ALR and HDAC4, can help us to identify thresholds with potential clinical utility.

ADIPONECTIN-LEPTIN RATIO

HDAC4

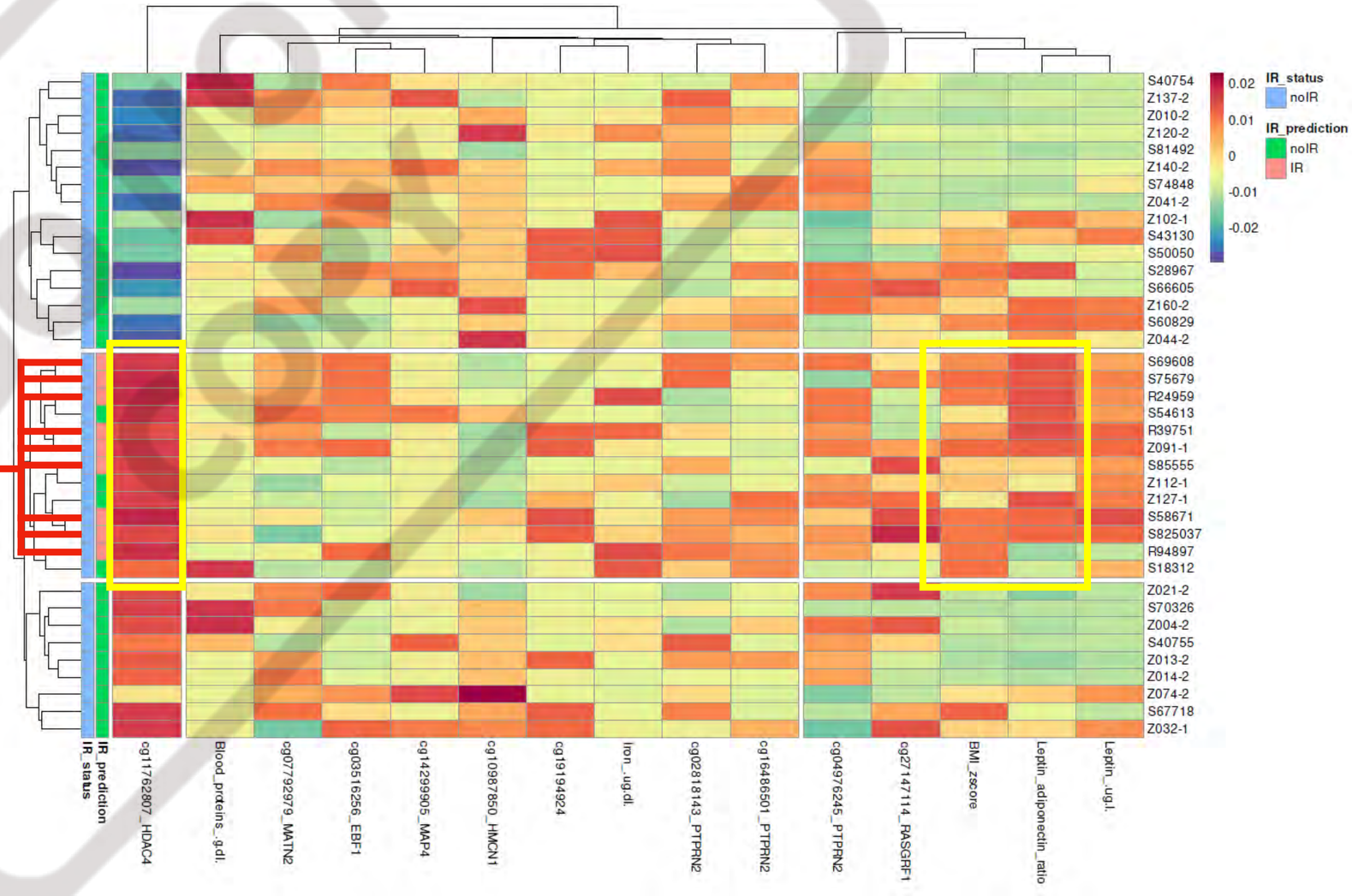
In the x axis the range of values for the predictor in each individual are shown

Novel ways of using SHAP values

USE SHAP FOR

Understanding model errors

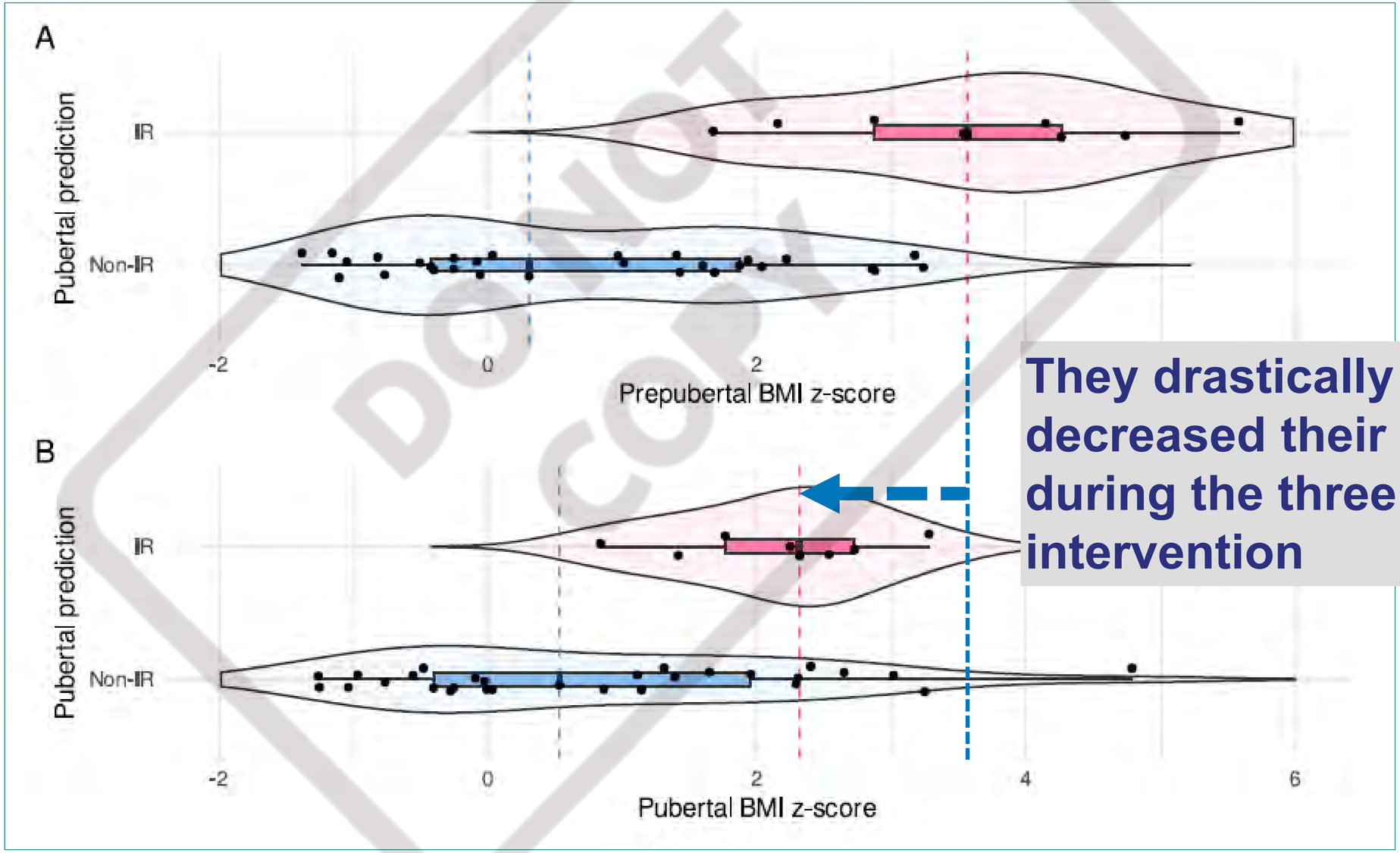
What happened in those individuals in which prediction failed?



Group of 9, which the model predicts as IR though they are noIR.

DNAm+ ALR&BMI+

What happens in these 9 subjects?

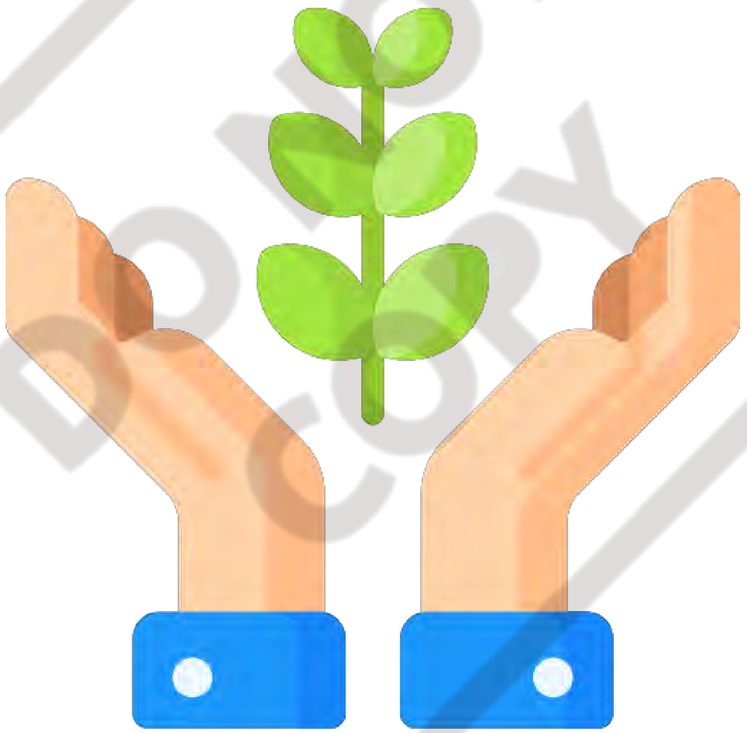


They drastically decreased their BMI during the three years intervention

What happens in these 9 subjects?

rastically
sed their BMI
the three years
ntion

A message of hope – predictions are not deterministic



KEY HOME MESSAGES

1. Specialists need to understand the decision-making mechanisms of ML-based CDSS, especially in high-risk areas such as healthcare, where system decisions may affect people's lives.

2. Integrating multi-omics and external data for predictive purposes require the careful design of analytic pipeline (data QUALITY, processing, fusion, model selection,…).

3. Transparent models should be prioritized for CDSS construction, and if not, post-hoc explainers should be used.

4. Post-hoc explainers such as SHAP can be exploited in many different ways (clustering, visualization techniques), to get the most out of predictive models.

MOOC AbiertaUGR

MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA. 5ª EDICIÓN

Inicio de matrícula: 10/02/2025

Inicio del MOOC: 10/03/2025

La Universidad de Granada pretende ofrecer un aprendizaje práctico y aplicado, accesible para todas las personas interesadas en el Machine Learning y Big Data para la Bioinformática. Para ello cuenta con un grupo de profesores de universidades, investigadores, profesionales y especialistas en cada una de las áreas, que ayudarán a introducirse en la Bioinformática y el Machine Learning en sus más amplios aspectos, aunando el rigor académico con [illegible]ncilla y directa que permita comprenderla y disfrutarla.

Acknowledgements

ISGlobal Instituto de Salud Global Barcelona

ciberobn isciii

ugr Universidad de Granada