Санкт-Петербургский Государственный Политехнический Университет

Институт прикладной математики и механики

КАФЕДРА
ТЕЛЕМАТИКА

**Семинар по специальности на английском языке**

**тема**

# Trust in AI: Model agnostic interpretation methods-

## separating the explanations from the (machine learning) model

**(Workshop in English)**

**Лекция 3**

---

16 Сентября
2020 г.

# Content :

- Main advantage of model-agnostic interpretation methods

(Основное преимущество методов интерпретации, не зависящих от модели)

- Why model-agnostic explanations method can be used for any type of model.

(Почему метод объяснения, независимый от модели, можно использовать для любого типа модели).

- High-Precision Model-Agnostic Explanations

(Объяснения, не зависящие от модели высокой точности

или

Объяснения высокой точности не зависящие от модели)

# Суть проблемы «моделей» сознания



Наше прошлое «записано» в нейросетях мозга, которые формируют то, как мы воспринимаем и ощущаем мир в целом и его конкретные объекты в частности.

Итого: в 99% случаев мы воспринимаем реальность не такой, какая она есть, а интерпретируем ее на основе готовых моделей (образцов) из прошлого.

# Metaphors for the issue under discussion

**Thinking is only calculation.**

(Мышление есть лишь расчет**)**

Томас Гоббс (1588-1679)

**Let's not argue - let's count**

(Не будем спорить — давайте посчитаем).

Жозеф Лагранж **(1736-1813)**

# AI as a issue of ..... computer science (CS)

## (ИИ как проблема... компьютерных наук)

**CS interpretation:**

AI problem is the solution to a new class of information processing problem, which includes 2 direct tasks and 1 invers task:

- Perception (восприятие) - obtaining data through communication or sensors channels

–> data - algorithm - data as an numbers

- understanding ( понимание) -matching the utility function to the data processing (согласование функции полезности с обработкой данных)

-> data - algorithm - data as an concepts

- knowledge itself (собственно познание) - building a logical model of perceived (воспринимаемых ) data and an algorithm for calculating a subset of the model state space on which utility function reaches maximum (алгоритм для вычисления подмножества пространства состояний модели, на котором функция полезности достигает максимума)

-> data – algorithm - algorithm

The such aspects composition can be considered as way to solve main AU inverse problem - “data-algorithm” . The essence of the task - “constructing an algorithm for computing a subset of the state space of models on which utility function achieves maximum .

# Main aspects of trust in formal explanation system

(Основные аспекты доверия к формальной системе объяснения)

**Model flexibility:** The interpretation method can work with flexible (different) type of AI models (random forests , deep neural networks…)

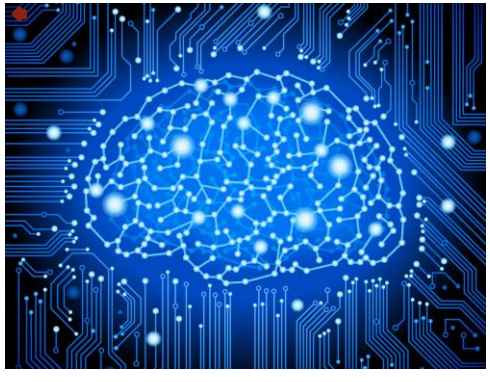
**Explanation flexibility:** There are several forms of explanation – simple linear formula, graphic, etc.

**Representation flexibility:** The explanation system should be able to use a different type of model representation as the model that previous being use for explained.

# Deal with the opacity (Разберитесь с непрозрачностью) - solving “inverse” problems in the space of “augmented reality”

Direct problem - прямая задача вычислений – моделирование объектов с помощью компьютеров с заданной архитектурой (АО+ПО):


Invers problem – выбор такого алгоритма вычисления, который «генерирует» данные, на которых функция полезности достигает максимума :



# High level look at model-agnostic interpretability:



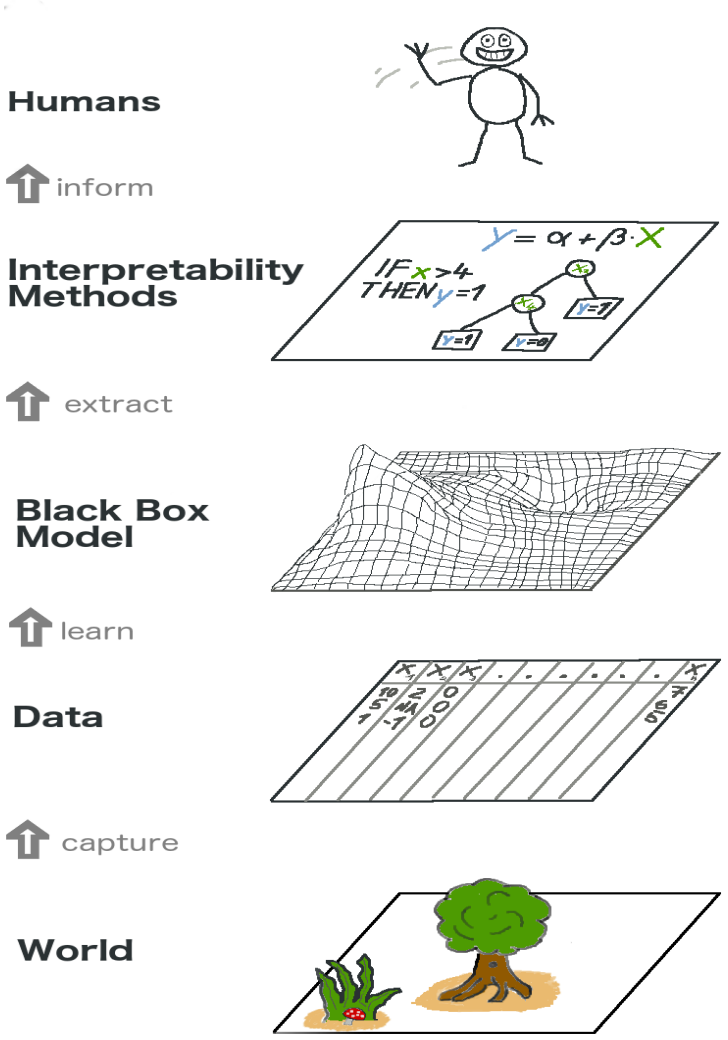
**Humans** - consumers of the explanations. Why we trust in …..

**Interpretability Methods** layer, which helps human deal with the opacity of machine learning models (how machine calculate explanations)

**Black Box Model** layer - algorithms using data from the real world to make predictions, find structures or invariants

**The Data layer** contains 'digital twins" anything from images, texts, tabular data and so on in order to make it process able for computers and also to store information.

**The World** layer contains everything that can be observed and is of interest.

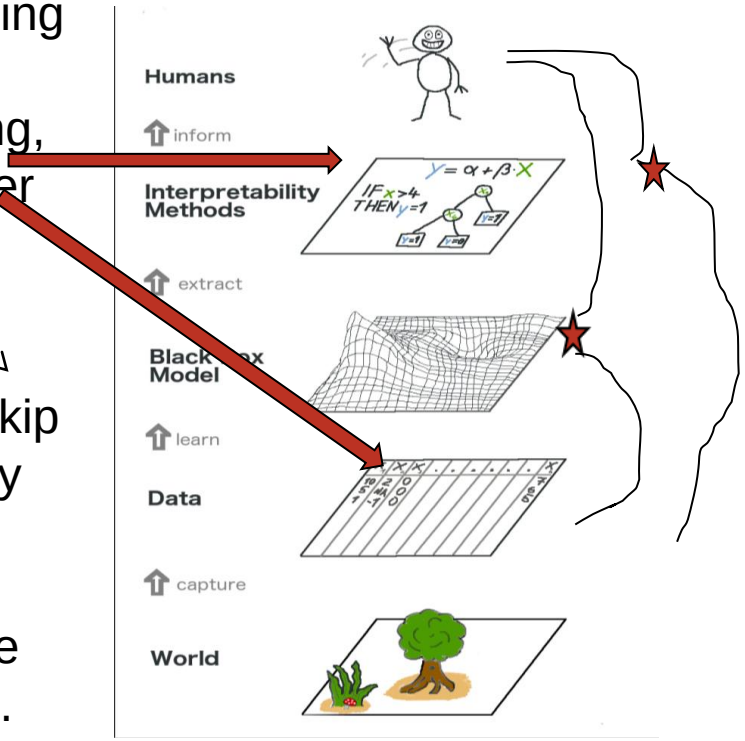


We need to understand the differences in approaches between statisticians and machine learning approaches.

Statisticians deal with the Data layer, such as planning, estimation, predictions, skip the Black Box Model layer and go right to the Interpretability Methods layer.

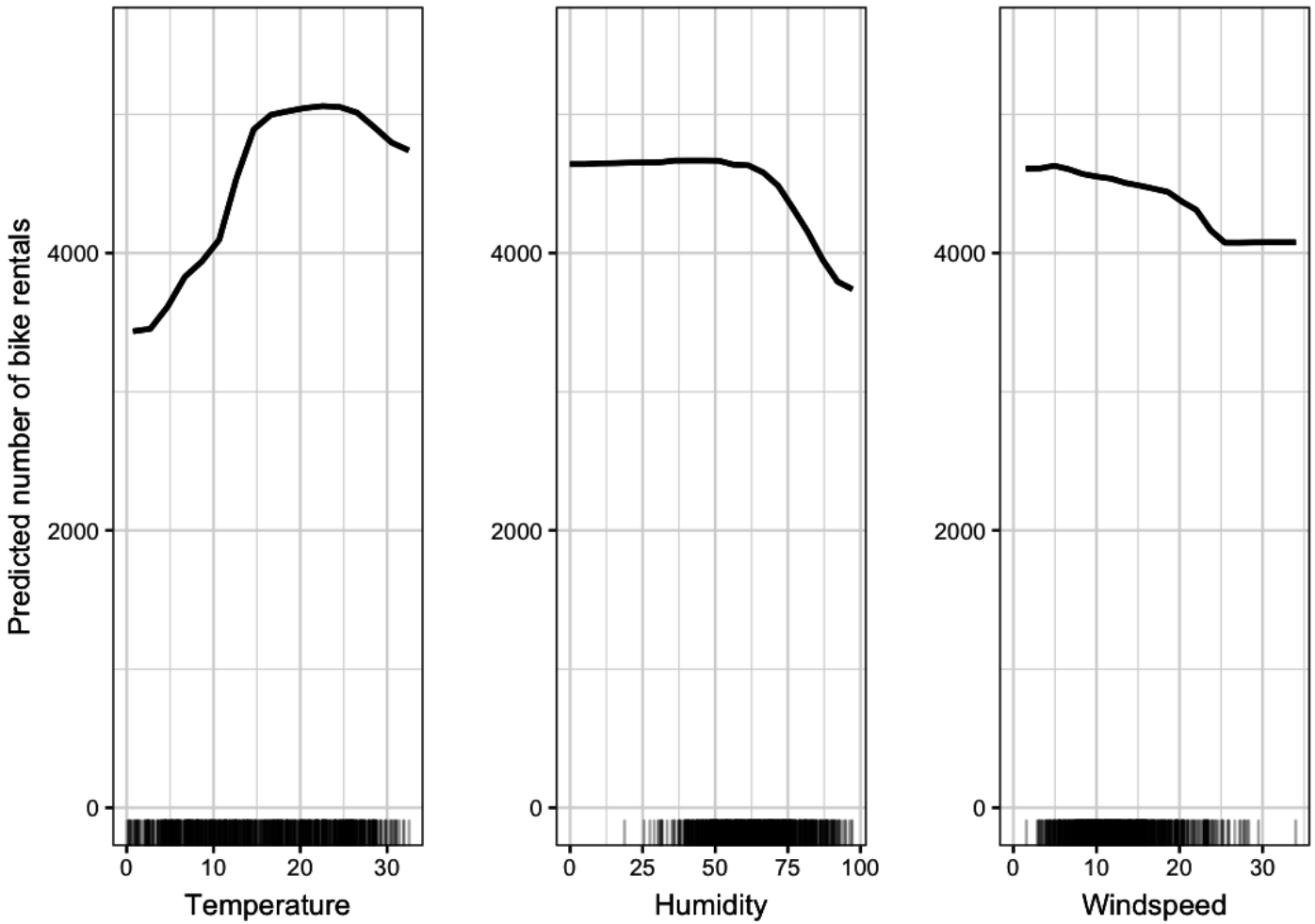
Machine learning specialists also deal with the Data layer, train a black box machine learning mode and skip The Interpretability Methods layer, so Humans directly deal with the Black Box model predictions.

Interpretable machine learning merge (unite, join) the work of statisticians and machine learning specialists.



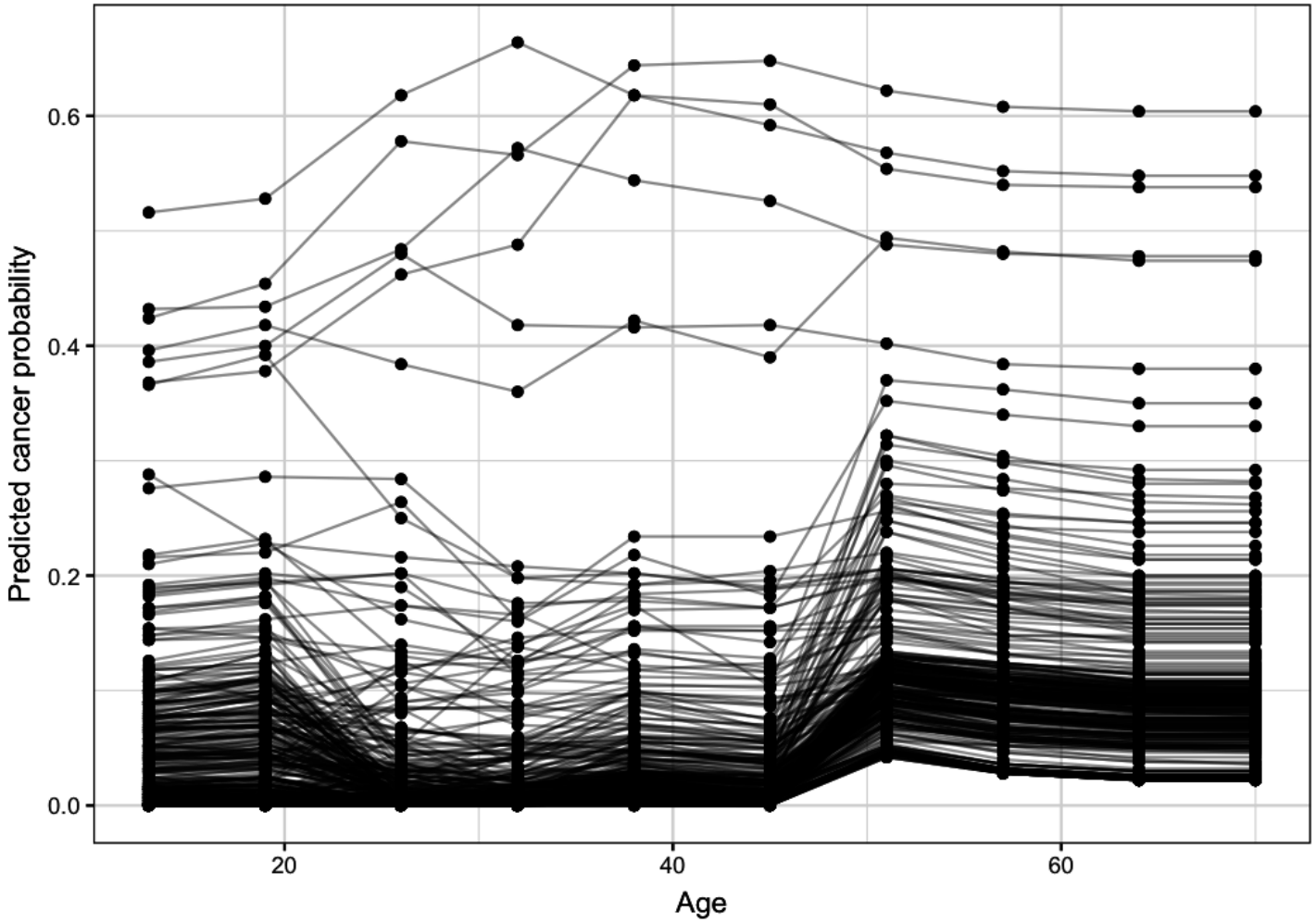
# Understanding is the search for "similar" to what has already been understood

как «работает» с числами сознание: «Все есть число» Пифагор



# From number to explanation

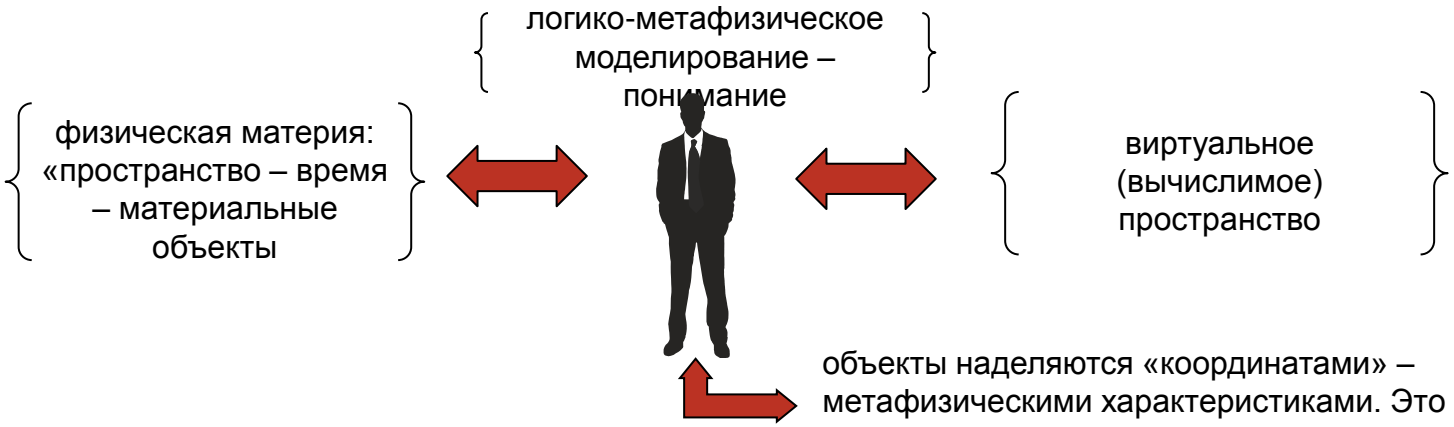




# Model-agnostic approach



## шаг 1




## шаг 2

На множестве объектов с координатами задается алгебра кардиналов множеств: операции сложения и умножения, вычисления характеристических функций – предикатов: ND->2D: $f(x_1, x_2, \dots, x_n) = \begin{cases} 1 \\ 0 \end{cases}$

# «Интеллектуальная» регуляризация



- Одна и та же система имеет различные физические свойства в зависимости от имеющейся информации (в одном случае она способна совершить работу, в другом – нет)
- Мера информации оказывается согласованной с общефизическими понятиями энергии и энтропии
- Информация как объективное описание состояния системы наравне с ее физическими параметрами меняет ее свойства. Т.е. в зависимости от имеющейся информации о системе систему можно или нельзя использовать для совершения работы. (в одном случае система способна совершить работу, в другом – нет)

# Новая «физика» .... мышления



Так, чтение или письмо – есть тренировка для головного мозга, в особенности если при этом вы узнаёте или выражаете нечто новое.

- Изменение сознания в процессе мышления приводит к изменениям в физическом теле интеллектуального субъекта.
- «Машина» обретет способность мыслить», если приобретет свойства «процессора управляемого данными»