YData

Accelerating AI
with improved data

YDATA.AI

AI is **not** delivering the expected results

YDATA.AI

85%

AI projects fail!

*Gartner

Why?

Limited access to data
due to privacy regulations and security policies

Poor data quality
is the biggest pain across all sectors

The **pains** of poor data quality

- Not enough data for relevant insights extraction

- Lack of labelled data by specialists

- Bias and inconsistencies

- Incomplete data due to missing values

- Underrepresented scenarios

- Hidden data due to privacy regulations

Data scientists have to deal with these challenges in a **non-standard** manner, while **lacking the time** and **tools** to achieve **optimal results**!

Data is a key factor for a successful AI adoption
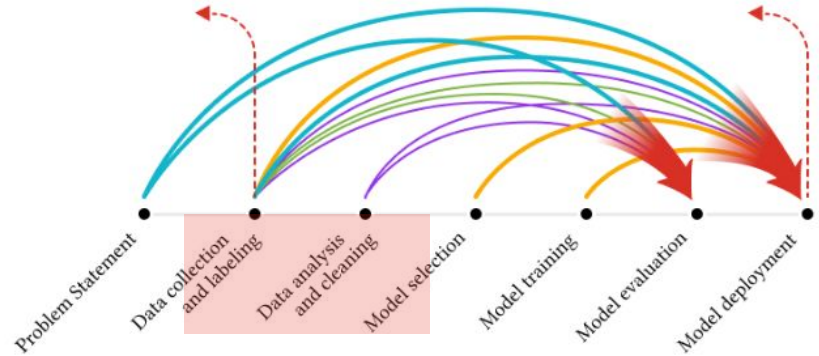
" (..) data quality issues in AI are addressed with the wrong tools created for, and fitted to other technology problems."

Data quality the most undervalued asset for successful AI
Google, 2021

The first data development platform

For data scientists, by data scientists.

YData's data development platform provides the necessary tools to evaluate and improve pre-training datasets, thereby reducing the costs while increasing the quality and efficiency of ML models.

Data discovery

Data connectors and
quality assessment

Data improvement

Synthetic data* and
data labelling

Experimentation at scale

Data pipelines and model training at scale

*Synthetic data is artificially generated data that was not collected from real world events. It keeps the same characteristics as the real data and is not traceable back to the original individuals.
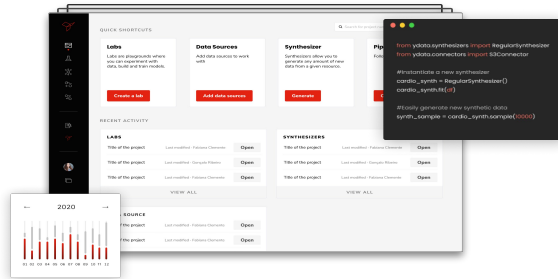
As simple as it can get for Data Science teams

Bring your own Data Sources

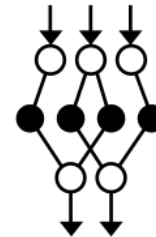## Collaborate & create

Achieve better results

+quality

+performance

YData

Labs

Playgrounds where data scientists can **build**, **mix** and **generate** datasets, **run** notebooks, **create** and **train** models at **scale**
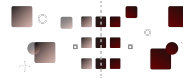

Data sources

Discover and **unlock** new data sources through **data connectors** and **exploratory data analysis** that includes metrics on the **metadata**, **inconsistencies** and **privacy**
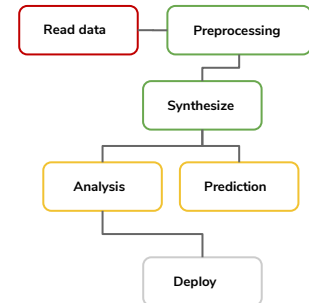

Synthesizers

Create **high quality** training data, while ensuring a **fast access** to data and **compliance** with **privacy regulations**

Pipelines

Simple and **easy scalability** of data flows by enabling users to build **ML pipelines** that **compare**, **fine-tune** and **contrast** multiple model's **metrics** and **results**

Synthetic data

A careful balance between utility and privacy

GDPR Anonymization	Risk-Based de-identification	Synthetic data
Removal of direct identifiers (e.g. names, IDs) while leaving other identifying info (birth data, zip-code, gender, etc)	Also called de-identification	Uses characteristics of the real data to generate new "fake" data.
Used for internal purposes.	Addresses risk related with Indirect identifiers	Models statistical distributions and structure of the real datasets.
Still considered personal information under some organizations.	No longer considered personal information.	Very similar to the original data in terms of granularity.
Additional safeguards and heavy protocols for access.	Anonymized data with potential loss of information-	Not considered personal information because data is not linked to actual activities.

How do we ensure synthetic **data quality**?

YData has an automated quality and privacy control process for every dataset generated with the goal to control the quality, utility, and privacy of the newly generated data.

For the quality, we use divergence metrics, correlation measures, and non-parametric tests, for the utility we apply the TSTR (Train Synthetic Test Real) methodology. In what concerns measuring the privacy leakage, we perform various tests, such as inference attacks.

Utility

98%

Privacy loss

3%
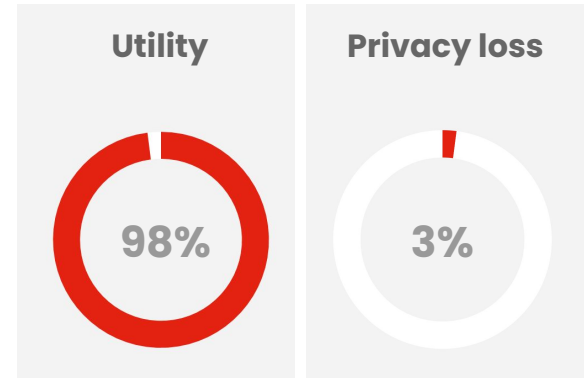
At the distance of a click or import statement

Data Science teams are now able to access data in **hours instead of months**, to leverage **bigger datasets instead of small** and not to worry with **compliance regulations** and **security** whenever there's the need to access new data

Training & Generating   Report

NAME	DATA SOURCE	CREATED
nameisready_TEST	Open	2020-12-10 14:10
● Ready to generate data		Retrain model

Data Generation History

Number of Records   **Generate**

CREATED	TOTAL RECORDS	ACTIONS
2020-12-10 14:10 demo@ydata.ai	10.000	Download Comparison report

```python
from ydata.synthesizers import RegularSynthesizer
from ydata.connectors import import S3Connector

#Instantiate a new synthesizer
cardio_synth = RegularSynthesizer()
cardio_synth.fit(df)

#Easily generate new synthetic data
synth_sample = cardio_synth.sample(20000)
```
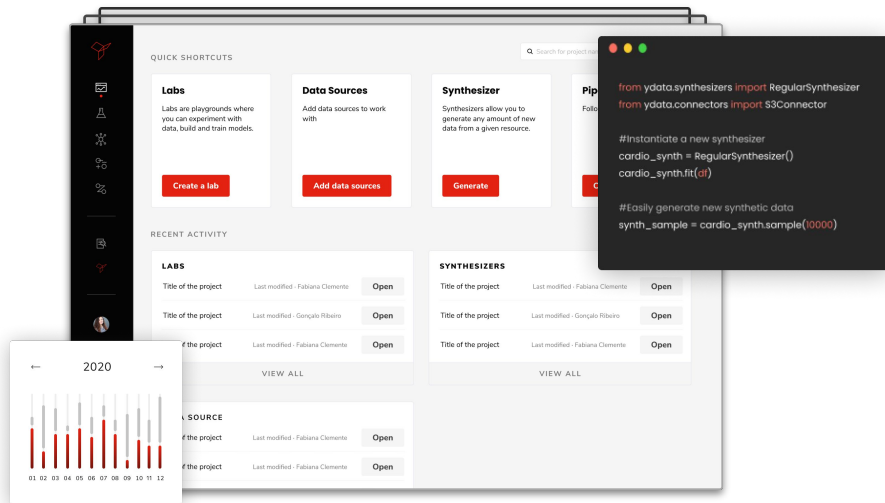
Cloud native with seamless integrations

One click deploy to every cloud or on-premises.
Highly scalable and available, connects to every database or data lake

No need for customization or implementation project

Integrates with other AI platforms

Adapts to every organization **authentication**

Easy management of projects and teams

YDATA.AI

Data
democratization

Break the silos across the
organization, while being
compliant with privacy
regulations.
Fast and easy access to new data
sources is a key driver for
innovation.

Higher RoI for
ML/AI

Data scientists time will be
invested in faster and better
development of solutions with
higher quality data, which is proven
to highly increase the RoI of
AI-based projects. Depending on
the use case, it can mean the twice
as much!

Creation of new
revenue streams

Becoming data-driven will allow
organizations to personalize
existing offerings and create new
products based on previously
locked data sources.
Monetization of data assets
becomes a reality.

Improve RoI for AI

Time and investment needed to deliver a model

Overhead

YDATA.AI

Higher productivity

Shorter roadmap

Higher revenue

Cost savings

Accelerating AI
with **improved data**

Early adopters of AI will be tomorrow's industry leaders.

Don't miss the opportunity, book a demo with us!

Gonçalo Martins Ribeiro

Chief Executive Officer

goncalo.ribeiro@ydata.ai