YData

Data-Centric AI in the era of LLMs

Data quality as the unfair advantage

by *Fabiana Clemente, CDO*

A bit about myself…

The AI Conference 2023

Fabiana Clemente, Chief Data Officer at YData

Applied Maths & Data Science

From big enterprises to startups

Data Science & Architecture

Co-Founder @YData

Interests

Data Science

Time-Series

Generative Models

LinkedIn

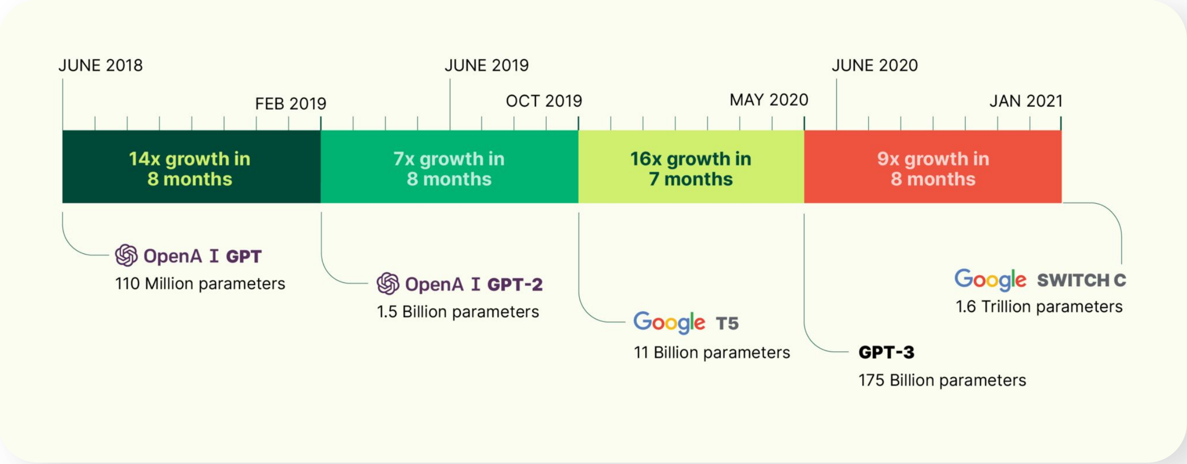
Foundation,

Foundation,

Foundation

The AI Conference 2023

Foundation Models: The future isn’t happening fast enough — Better tooling will make it happen faster



JUNE 2018

FEB 2019

JUNE 2019

OCT 2019

MAY 2020

JUNE 2020

JAN 2021

14x growth in 8 months

7x growth in 8 months

16x growth in 7 months

9x growth in 8 months

OpenA I **GPT**
110 Million parameters

OpenA I **GPT-2**
1.5 Billion parameters

Google **T5**
11 Billion parameters

GPT-3
175 Billion parameters

Google **SWITCH C**
1.6 Trillion parameters

The AI Conference 2023

Data as a product!

Thinking data as a product means putting the business needs at the heart of the data flows/preparation design.

... prioritize the quality of your data!

The dimensions of data quality

From raw to smart data

Accuracy

Completeness

Consistency

Integrity

Validity

Uniqueness

*Dimensions of Data Quality

- Clarity
- Availability
- Accuracy
- Comparability over time
- Compliance with laws and regulations
- Granularity
- Interpretability
- Relevance
- Variety
- …

The in(complete) data

Impact of missing data in time-series of different verticals

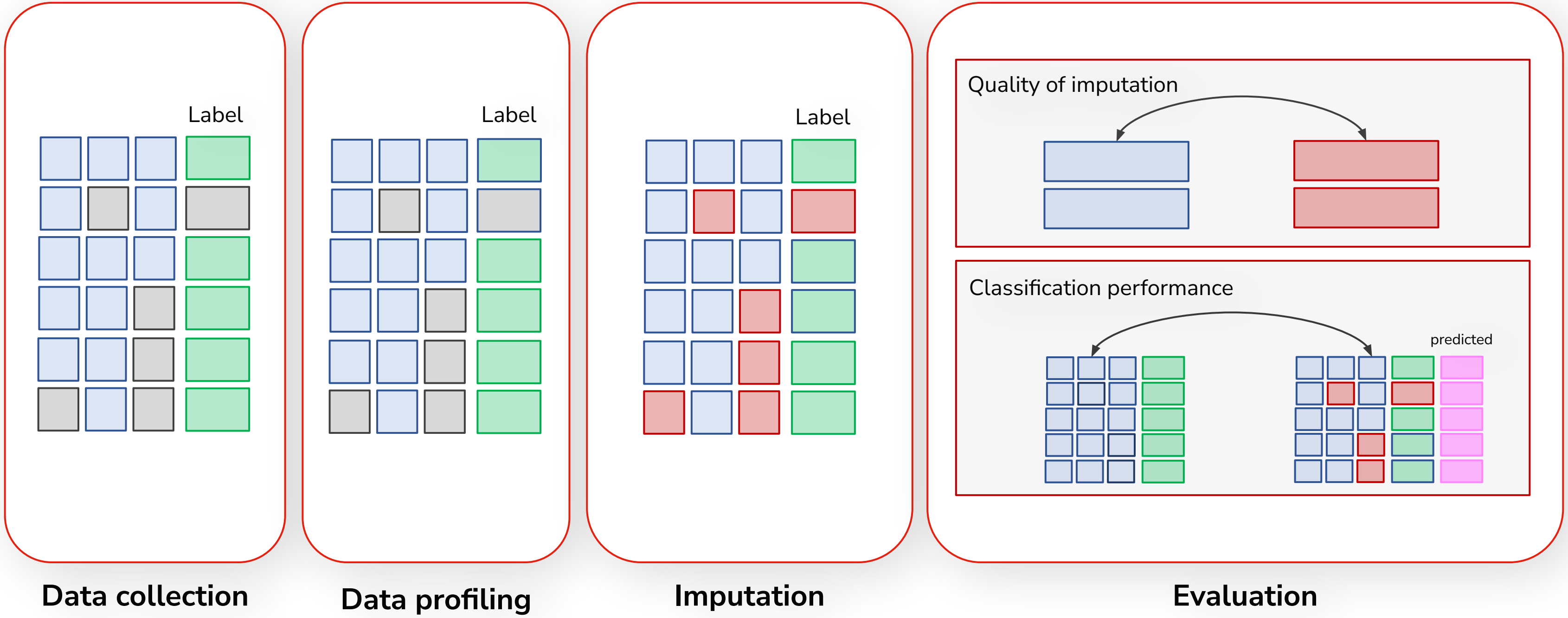
Inaccurate risk management: The presence of missing data can lead to underestimation or overestimation of risks.

Hindered quality control: can lead to a poor identification of defects or deviations in production processes.

Medical misdiagnosis: patient care can be negatively impacted due to lack of good disease tracking information.

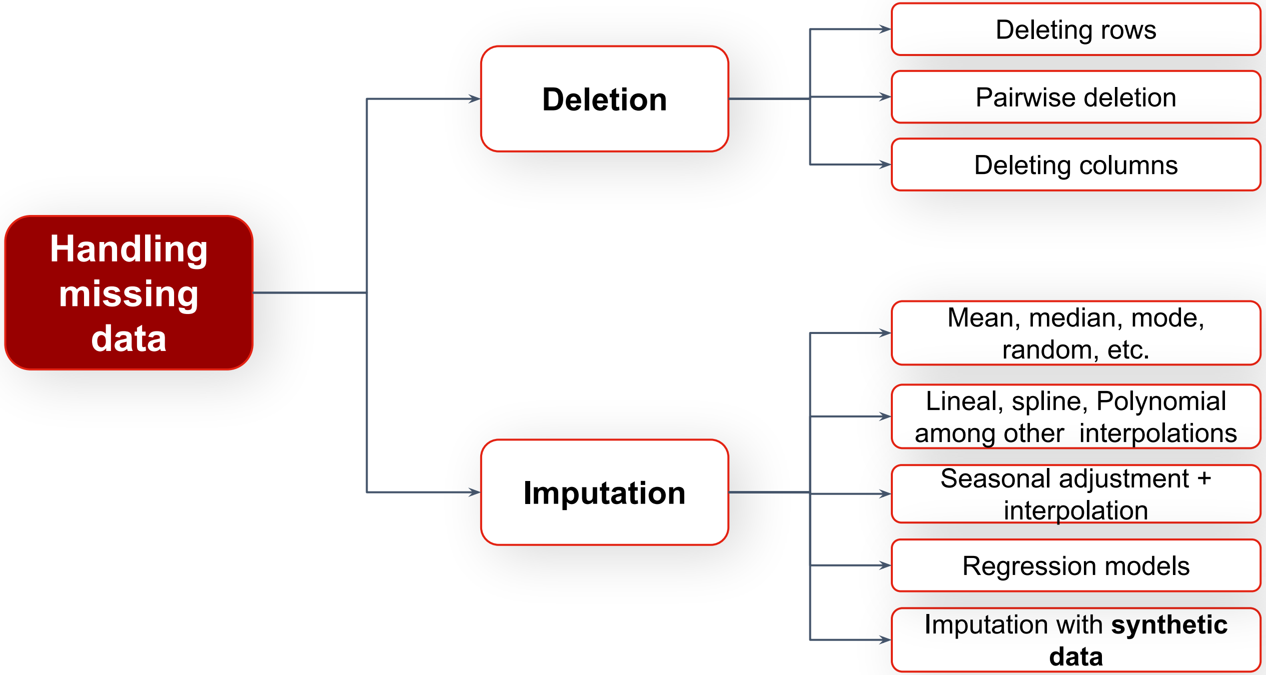
Optimized missing data imputation

The missing data workflow for time-series



Time-series missing data imputation

Not all the methods have the same result!



Handling missing data

Deletion

Deleting rows

Pairwise deletion

Deleting columns

Imputation

Mean, median, mode, random, etc.

Lineal, spline, Polynomial among other interpolations

Seasonal adjustment + interpolation

Regression models

Imputation with **synthetic data**

The AI Conference 2023

Generative models for missing data

A data synthetic data generation approach for time-series imputation

- Synthetic data is artificially generated data that was not collected from real world events.
- The generated data is the result of the learning of underlying multivariate data distribution conditioned to the behaviour that we want to predict;
- Generative models are more flexible being able to adapt to short and long-term gaps;

GANs

Transformers

VAEs

Bayesian Net

GMMNs

Markov chains

Why a synthetic data approach?

A solution to overcome traditional imputation methods gaps!

The AI Conference 2023

Data anomaly handling

SD generation can produce data points that are less sucesptible to anomalies or outliers.

Flexibility & adaptability

Generative models can be tailored to specific data types and structures, making it easy to adapt.

Reduced bias & variance

High-quality SD can reduce bias and variance in analysis results, as imputed values are generated with better context.

Speed & scale

Generative models can be trained on large datasets efficiently, often leveraging parallel processing and GPU acceleration.

Use case: Missing data in stock data

The impact of different missing data gaps in time-series

The AI Conference 2023

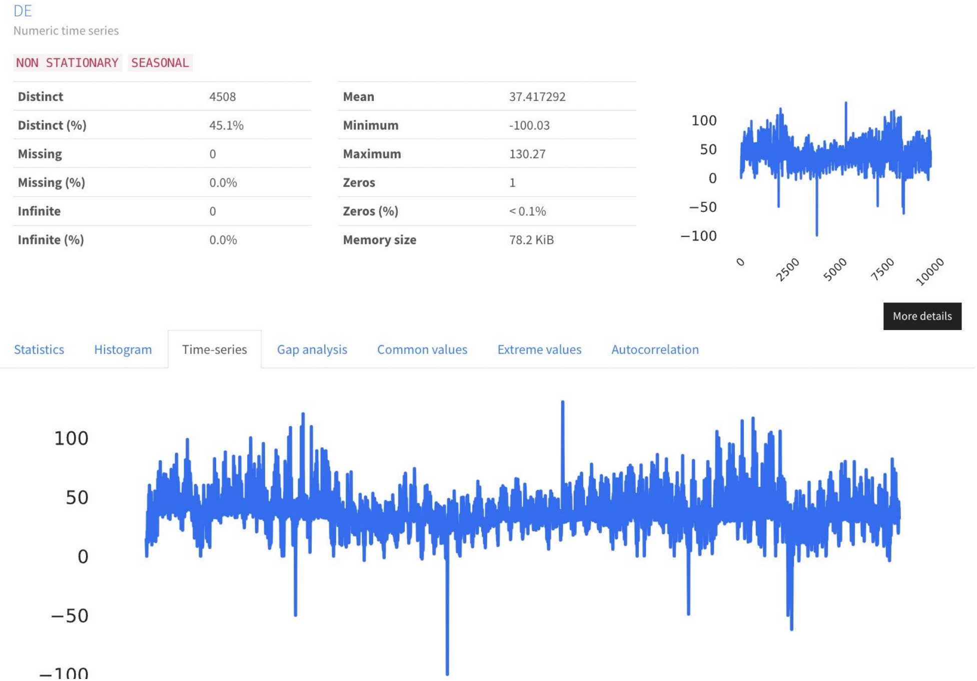
Dataset

The stock dataset includes the stock information for 4 different companies for the period between April 2013 and April 2014. The socks include Deere, First Industrial and

Dataset characteristics & challenges

Variables are non-stationary time-series with high seasonality patterns.

High-correlation between some of the variables.



DE

Numeric time series

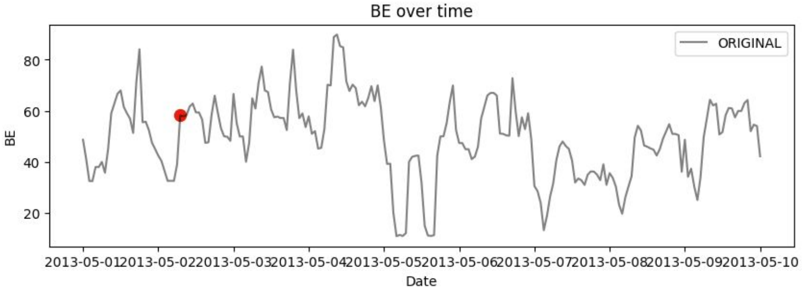
NON STATIONARY SEASONAL

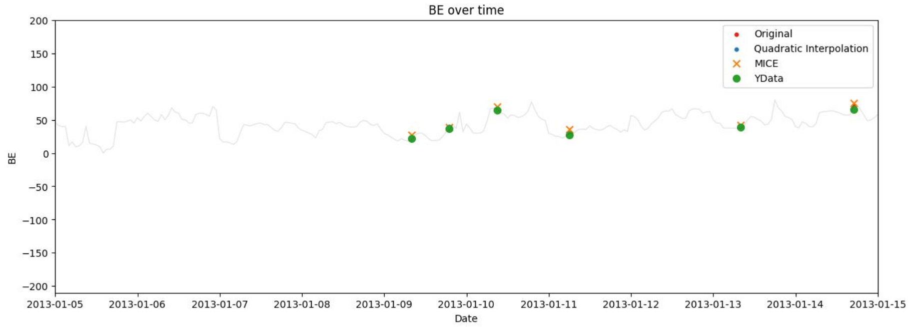
Distinct	4508	Mean	37.417292
Distinct (%)	45.1%	Minimum	-100.03
Missing	0	Maximum	130.27
Missing (%)	0.0%	Zeros	1
Infinite	0	Zeros (%)	< 0.1%
Infinite (%)	0.0%	Memory size	78.2 KiB

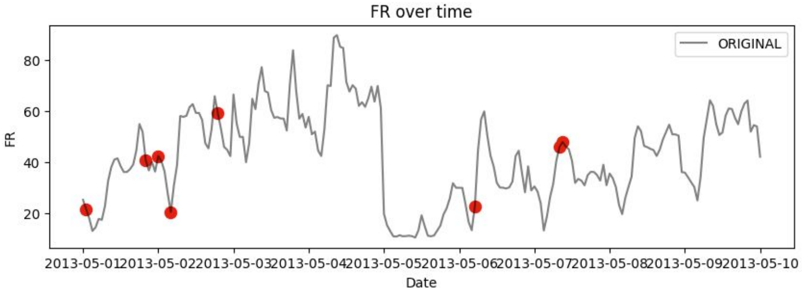
Small gaps

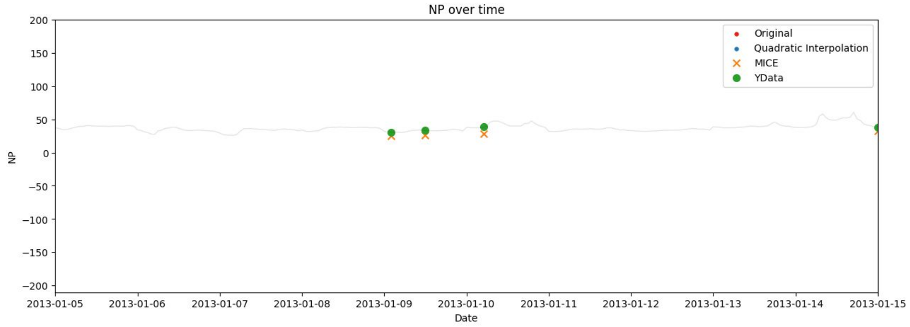
What if we have only a few hours to fill-in?

The AI Conference 2023


BE over time

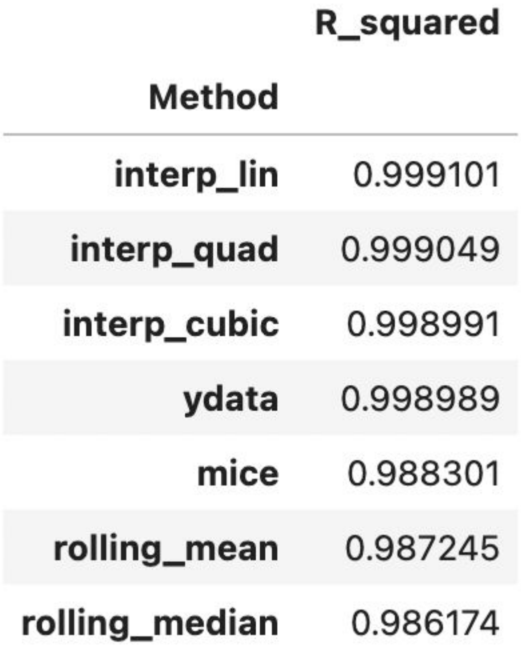

BE over time


FR over time

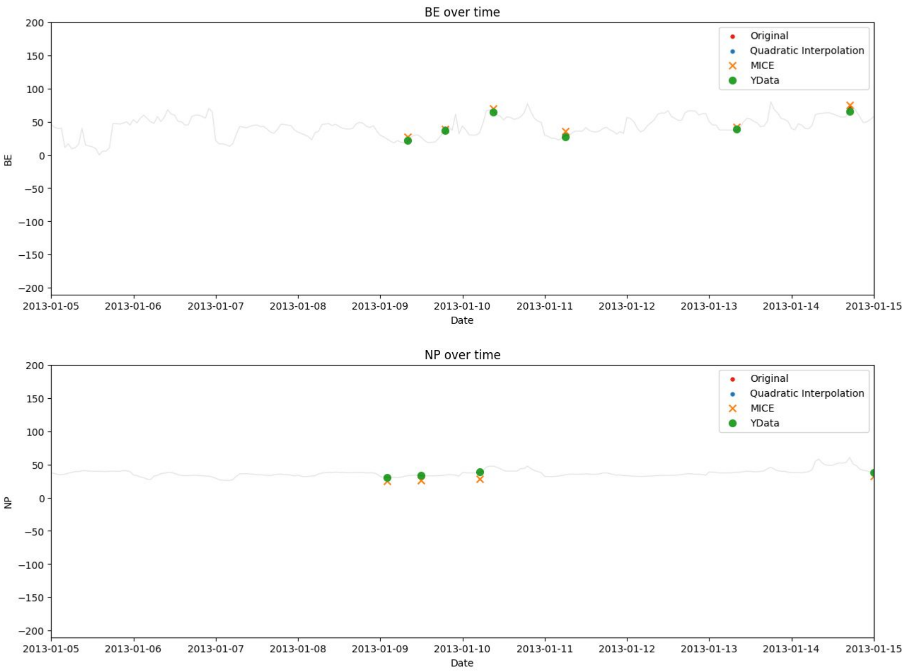

NP over time

Small gaps

What if we have only a one or 2 hours to fill-in?

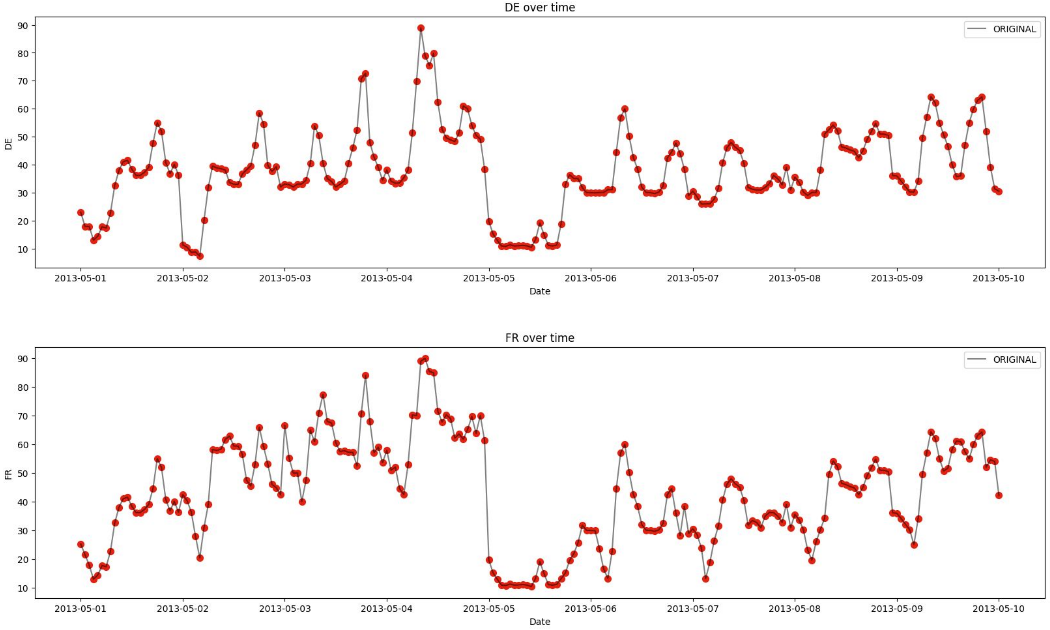


	R_squared
Method	
interp_lin	0.999101
interp_quad	0.999049
interp_cubic	0.998991
ydata	0.998989
mice	0.988301
rolling_mean	0.987245
rolling_median	0.986174



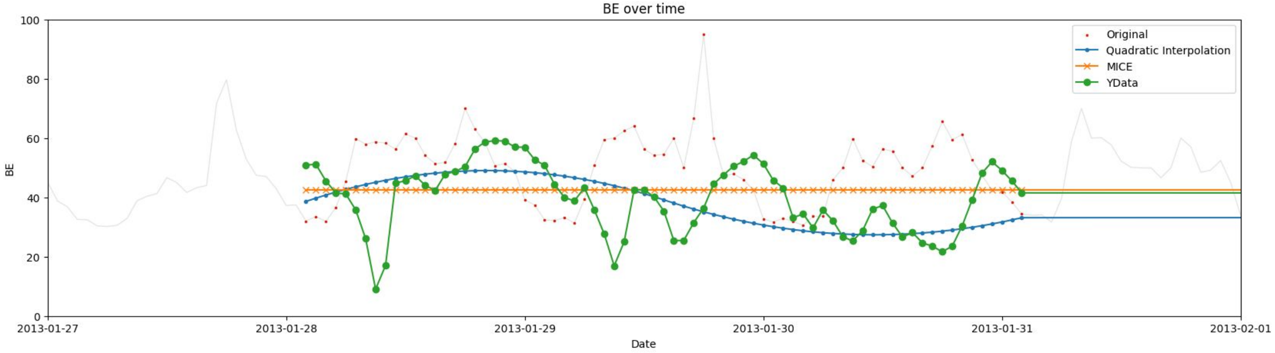
Medium gaps in stock data

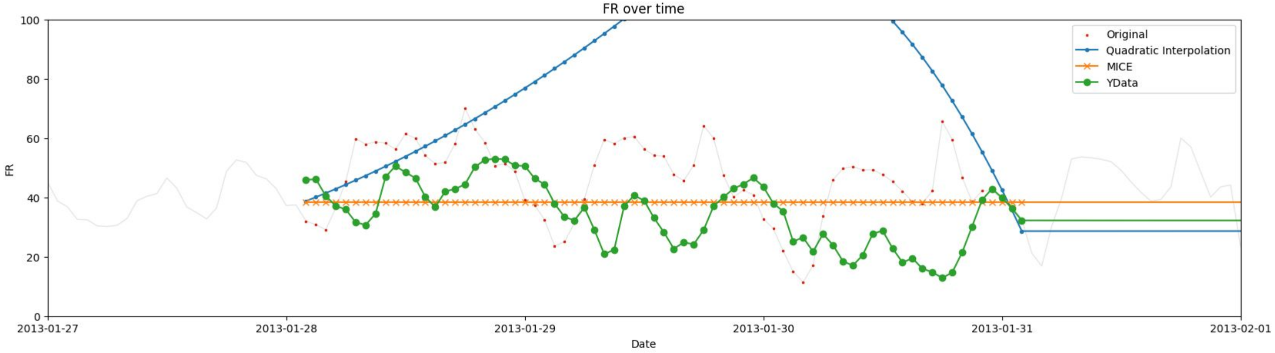
What if we have only a few hours to fill-in?

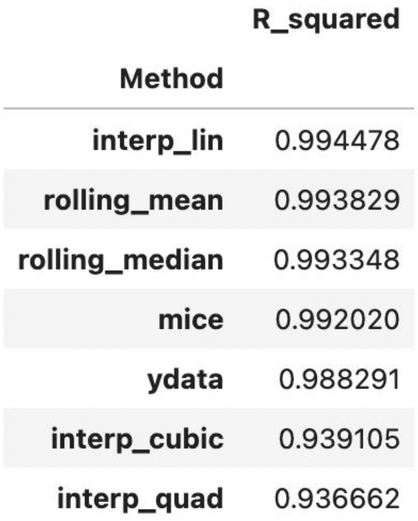


Medium gaps in stock data

What if we have only a few hours to fill-in?



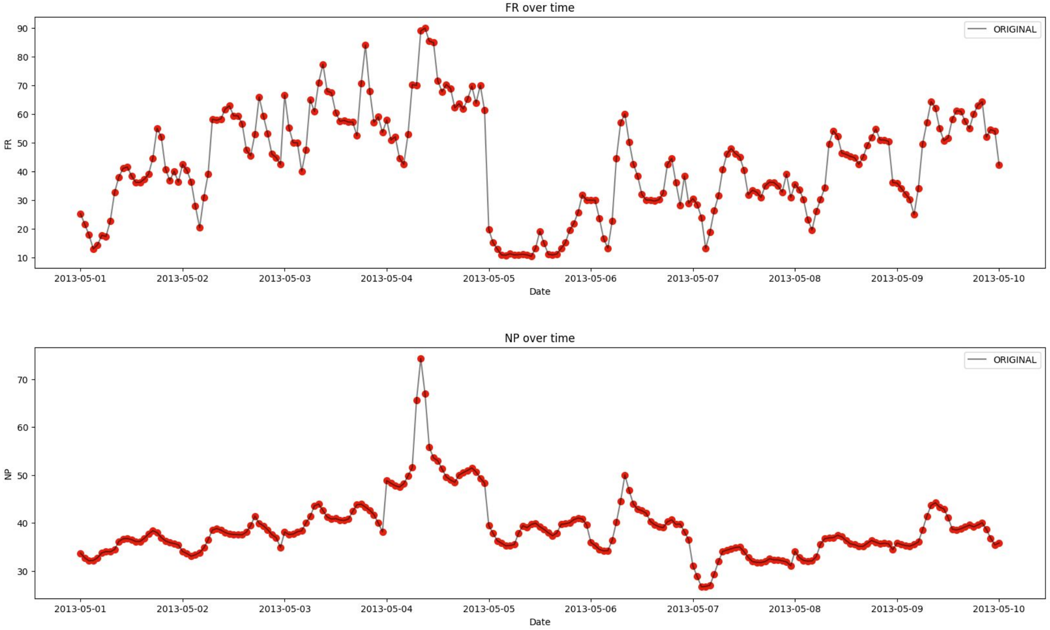




	R_squared
Method	
interp_lin	0.994478
rolling_mean	0.993829
rolling_median	0.993348
mice	0.992020
ydata	0.988291
interp_cubic	0.939105
interp_quad	0.936662

Long gaps in stock data

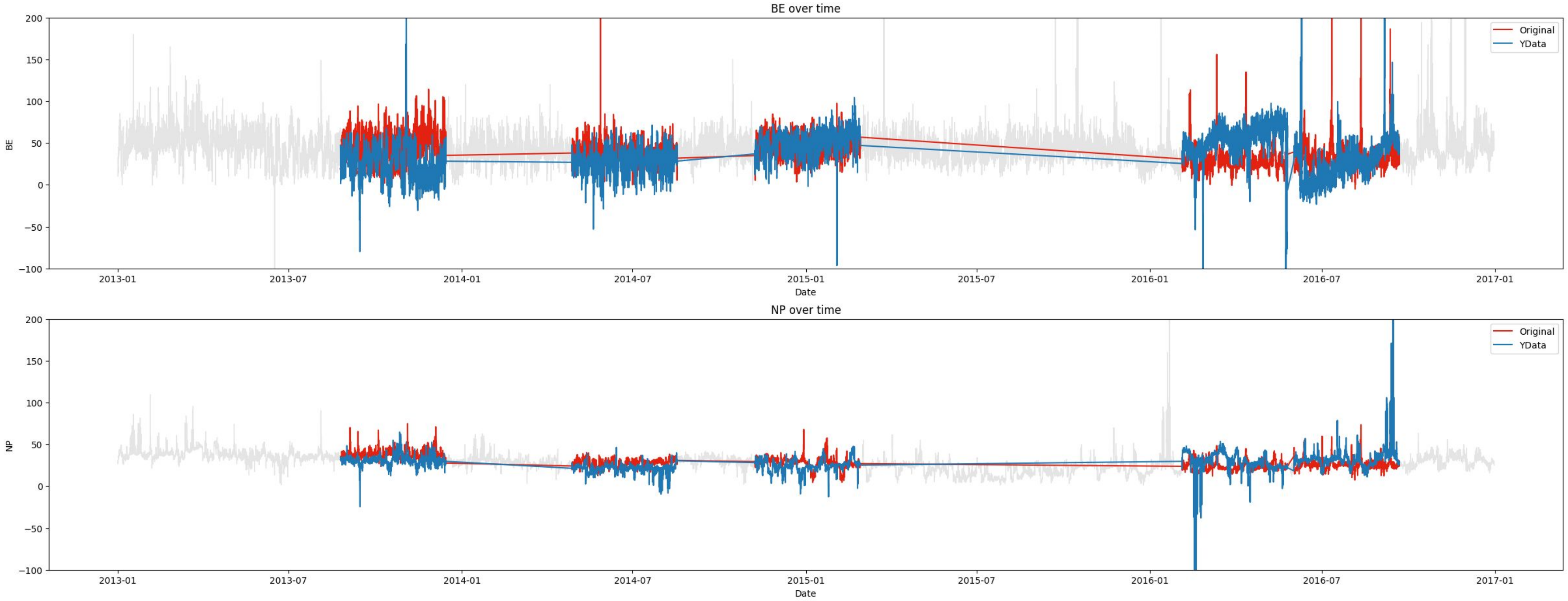
What if we have only a several days in a row to fill-in?



The AI Conference 2023

Long gaps in stock data

What if we have only a several days in a row to fill-in?



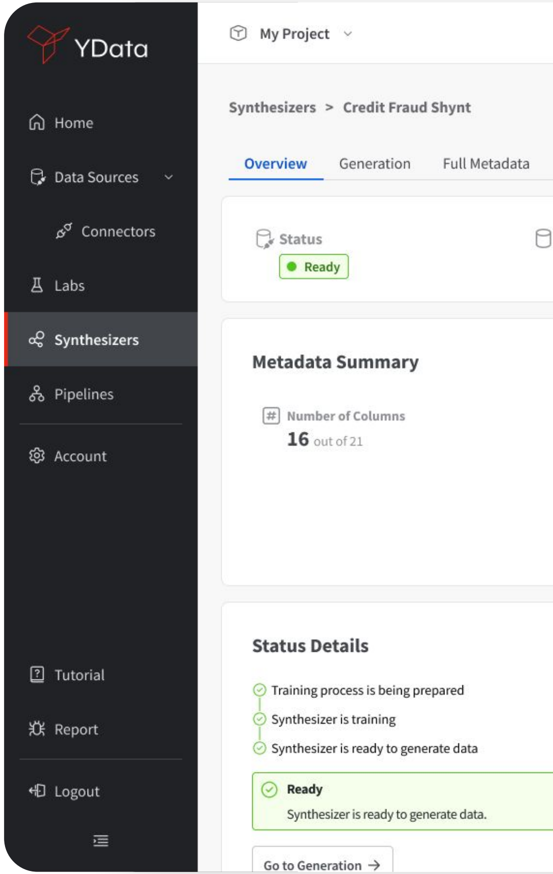
Synthetic data to fill-in the gaps

What you should recall!

1. Comes in many forms
2. Is able to capture the complexity of many different distributions and datasets
3. Can be used to address data quality issues: such as missing data imputation
4. Generative models & Synthetic data are able to cope with challenging and long-periods of missing gaps.
5. It is cost-efficient and can remove access barriers

Register at **ydata.ai/register** to try it out!

PyData Seattle 2023



Training process is being prepared

Synthesizer is training

Synthesizer is ready to generate data

Ready

Synthesizer is ready to generate data.

Accelerating AI with improved data

LLMs and Generative Revolution

> “The problem YData is solving is foundational and core to machine learning. It is know that the quality of data is the most important asset for an AI solution and ensuring it is something really hard and expensive.”

Paul Horn, ex-SVP & Director of Research @IBM, DIstinguished Scientist @NYU

Fabiana Clemente, *CDO*

fabiana.clemente@ydata.ai