Getting From Data to Insights With Powerful XAI & DataViz Open-Source Tool

Raphaël Lüthi

PyData London 2023

4th June 2023

Who am I ?

Raphaël Lüthi

Machine Learning & AI Lead @ Groupe Mutuel

LinkedIn

2010 — EPFL — **Engineering MSE**

2016 — Totaljobs — **Data Scientist**

2019 — SBB CFF FFS — **Data Scientist**

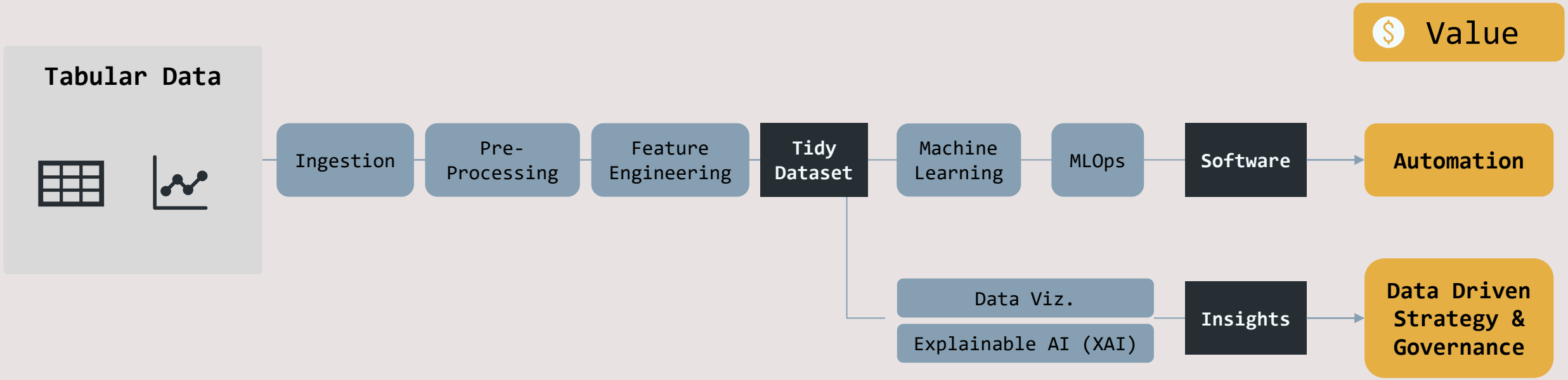
2020 — groupemutuel — **Machine Learning & AI Lead**

Today

Agenda

1. **Into : How to Create Value as a Data Scientist?**
2. The SHAP Explainable AI Method
3. Demo: My Workflow on the California Housing Dataset
4. Successful Applications in a Business Setting

How to Create Value as a Data Scientist?



Agenda

1. Into : How to Create Value as a Data Scientist?
2. **The SHAP Explainable AI Method**
3. Demo: My Workflow on the California Housing Dataset
4. Successful Applications in a Business Setting

Data to Insights in Theory

Using Shap Values to Extract Insights From Data as Proposed by Lundberg et al. 2020

Explainable AI for Trees: From Local Explanations to Global Understanding

Scott M. Lundberg[1], Gabriel Erion[1,2], Hugh Chen[1], Alex DeGrave[1,2], Jordan M. Prutkin[3], Bala Nair[4,5], Ronit Katz[6], Jonathan Himmelfarb[6], Nisha Bansal[6], and Su-In Lee[1,*]

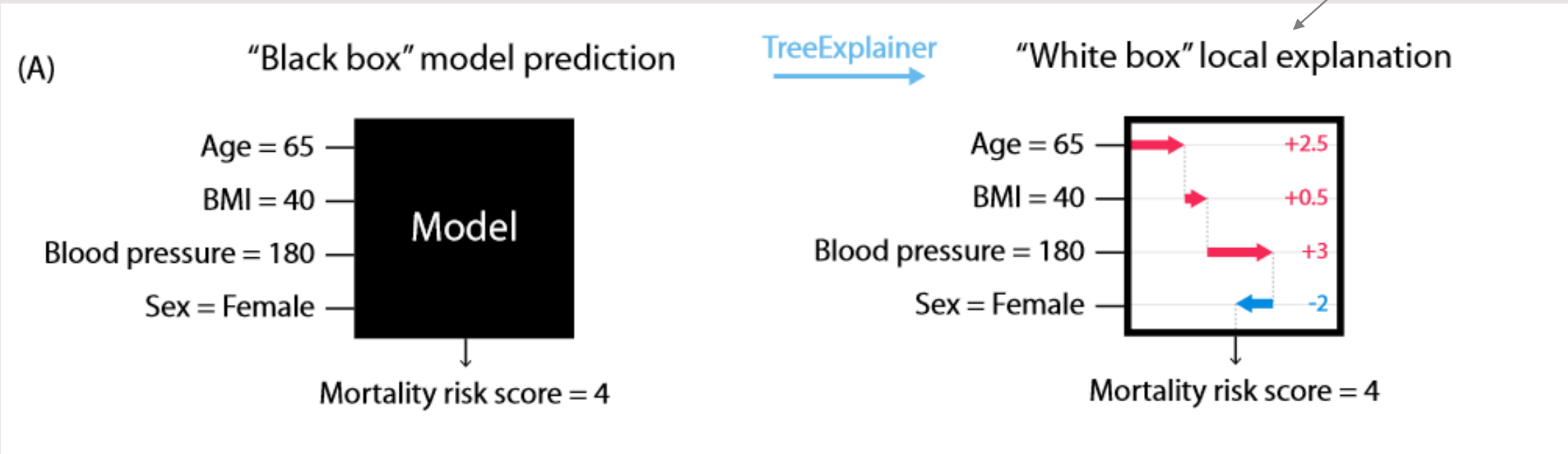
Shapley Additive exPlanations

Sources: https://www.nature.com/articles/s42256-019-0138-9; https://shap.readthedocs.io/en/latest/index.html

Local Explanation with SHAP Values

Lundberg et al. 2020 :

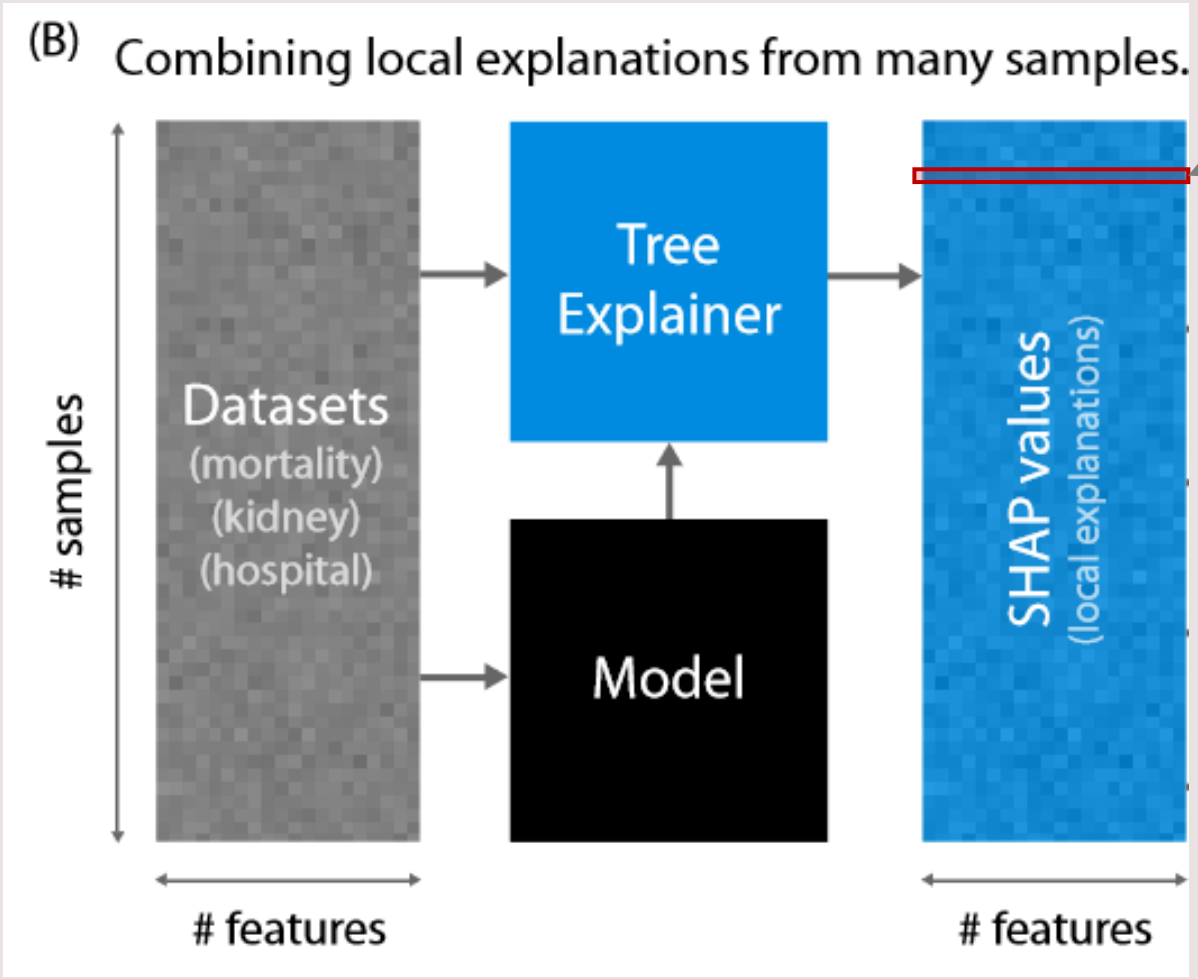
Explainable AI (XAI)

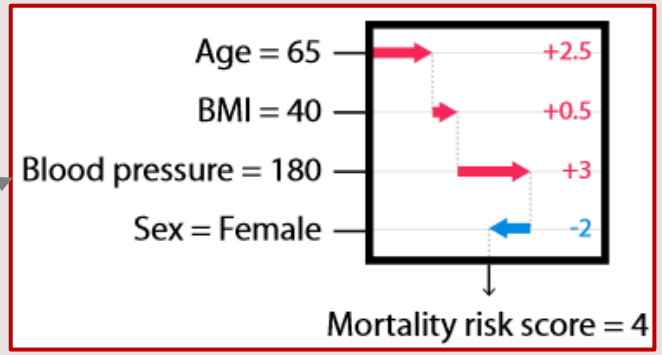


Source: https://www.nature.com/articles/s42256-019-0138-9

Global Explanation with SHAP Values

Lundberg et al. 2020 :

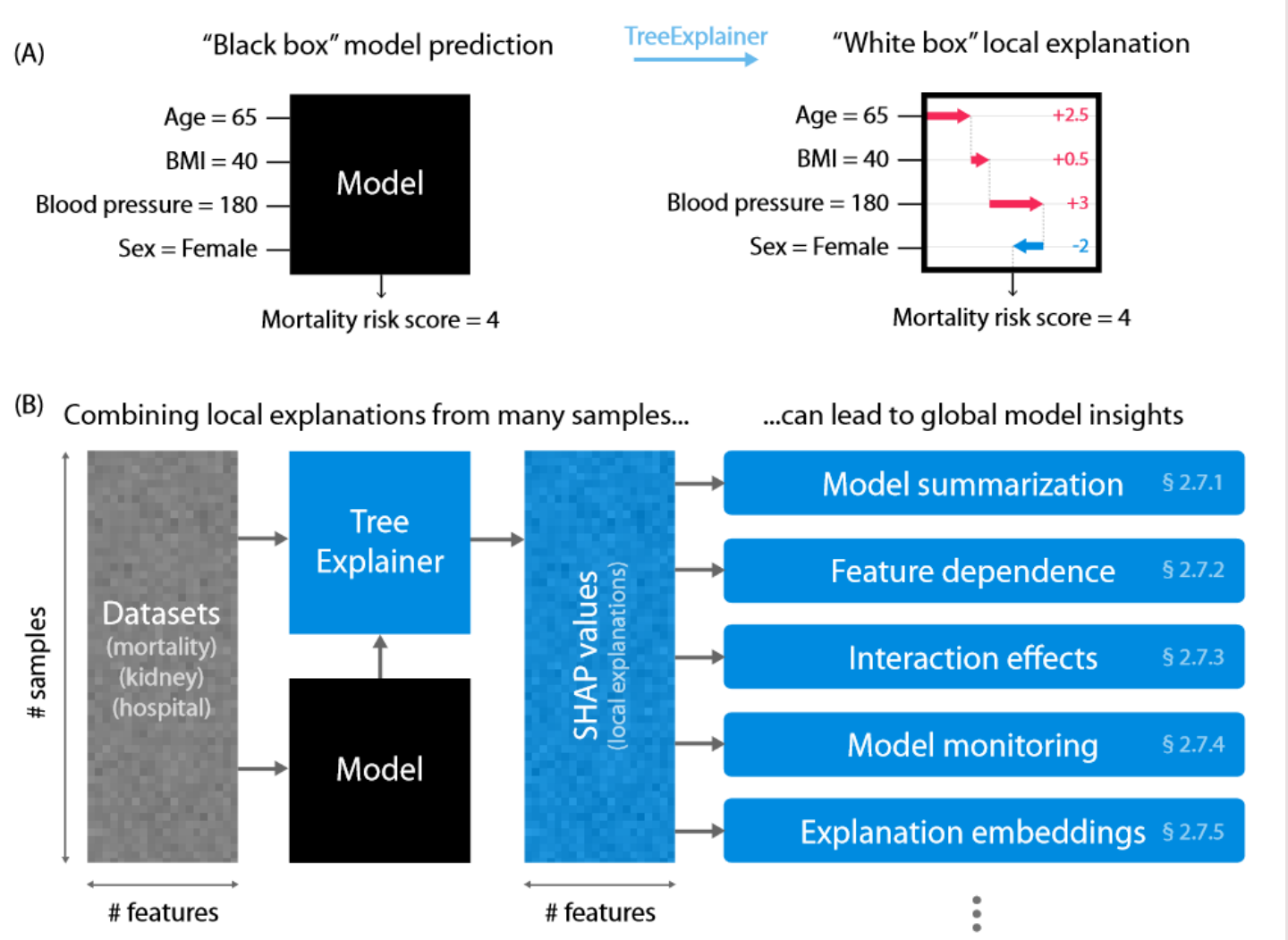




Source: https://www.nature.com/articles/s42256-019-0138-9

Global Insights with SHAP Values

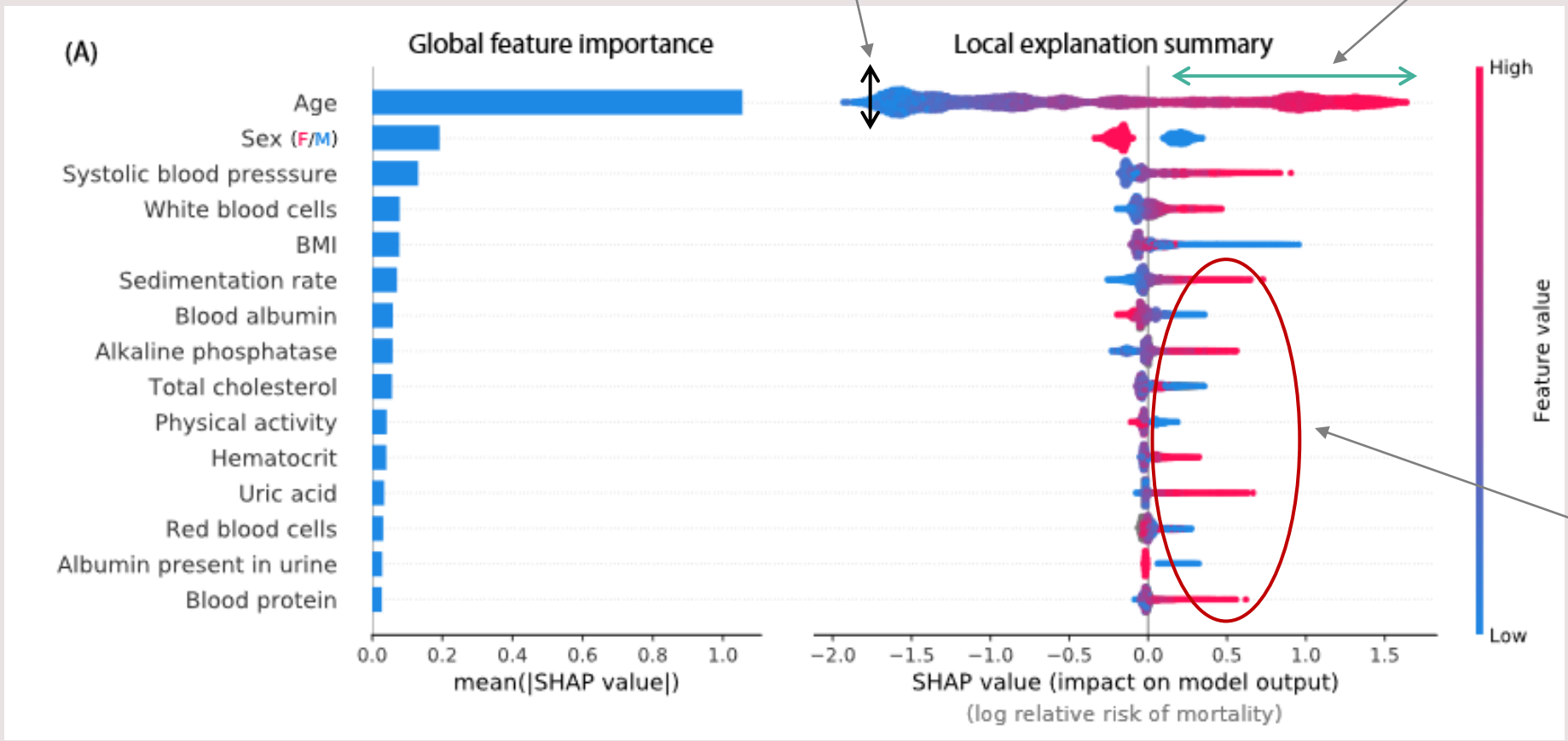
Lundberg et al. 2020 :



Source: https://www.nature.com/articles/s42256-019-0138-9

SHAP Summary Plots

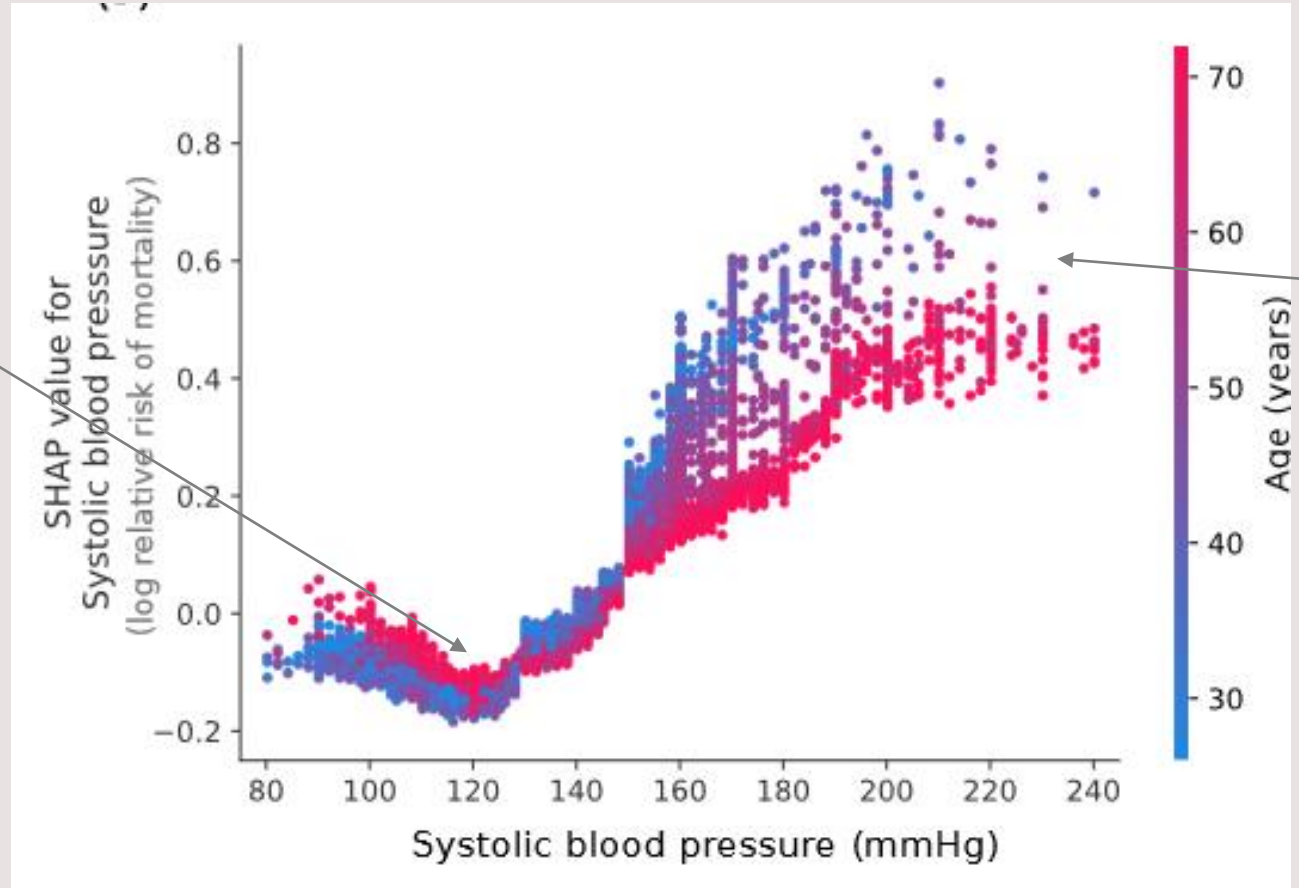
Lundberg et al. 2020 :



Source: https://www.nature.com/articles/s42256-019-0138-9

SHAP Dependence Plot

Lundberg et al. 2020 :



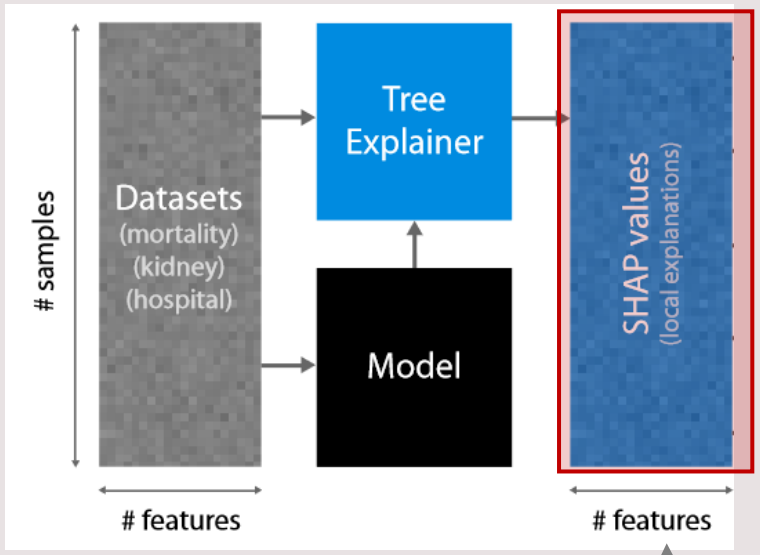
Threshold

Complex Nonlinear Interactions

Source: https://www.nature.com/articles/s42256-019-0138-9

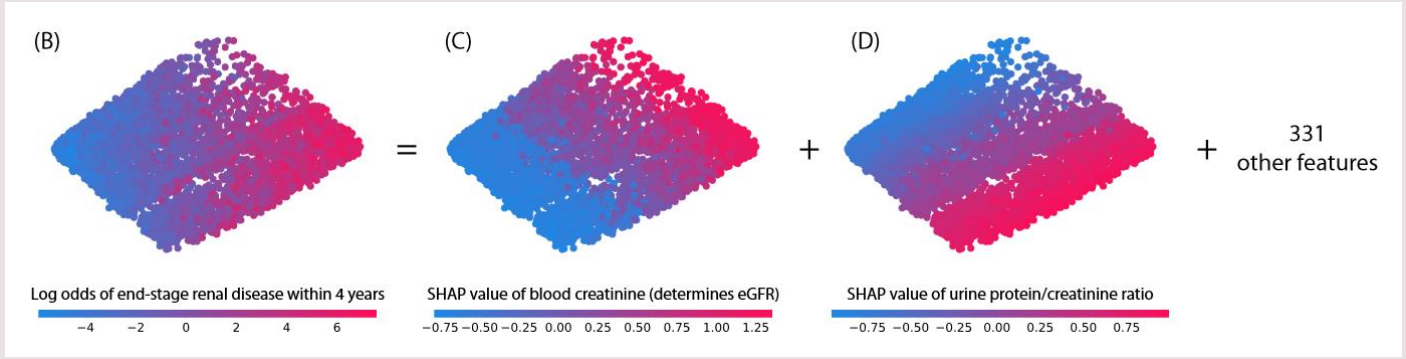
SHAP Explanation Embedding and Clustering

Lundberg et al. 2020 :

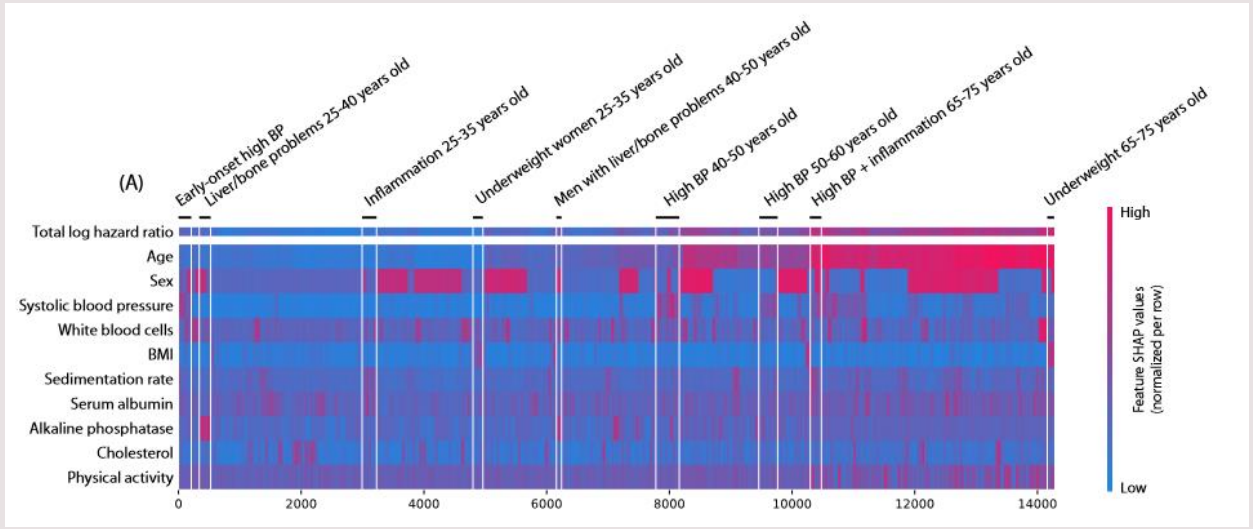


Matrix of Float Values

Explanation Embedding



Explanation Clustering (semi-supervised)



Source: https://www.nature.com/articles/s42256-019-0138-9

Agenda

1. Into : How to Create Value as a Data Scientist?
2. The SHAP Explainable AI Method
3. **Demo: My Workflow on the California Housing Dataset**
4. Successful Applications in a Business Setting

Follow Along the DEMO :

My Workflow on the California Housing Dataset

GitHub

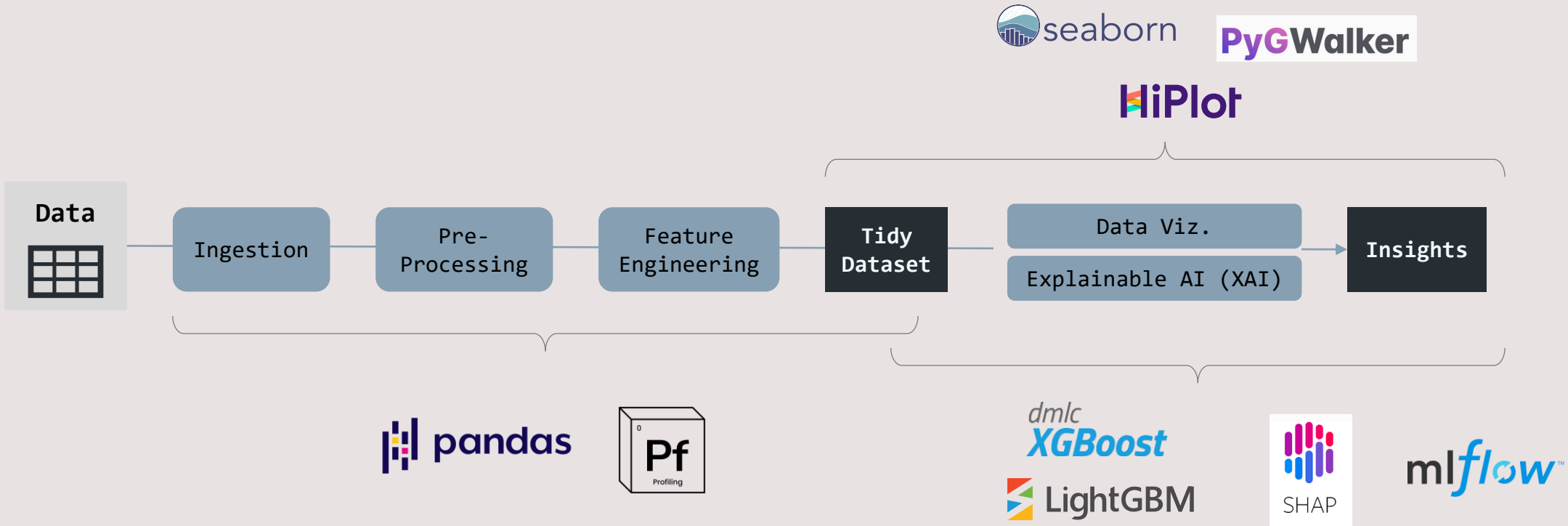
https://github.com/rluthi/pydata-london-2023

Google colab

https://colab.research.google.com/github/rluthi/pydata-london-2023/blob/main/01_Data-to-Insights-Demo_PyData-London-23.ipynb

Data to Insights in Practice

My Battle Tested Workflow Using Open-Source Tools



A Calm Way to Organise Your Data Wrangling

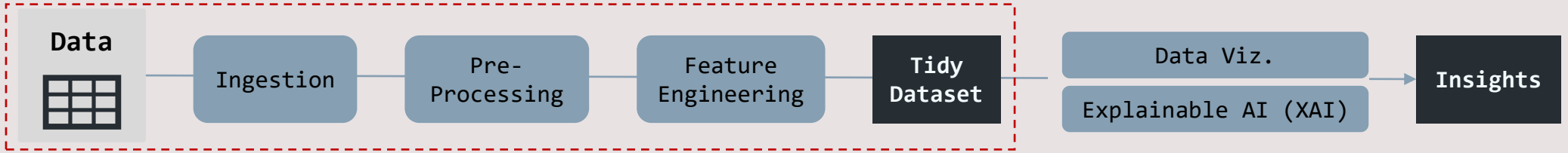
pandas pipe

pandas pipe

Full tutorial: calmcode.io/pandas-pipe/introduction.html

```python
data = (
    data_raw.pipe(start_pipeline)
    .pipe(query, query="MedInc <= 10")
    .pipe(query, query="AveRooms <=10")
    .pipe(query, query=("0.8 <= AveBedrms <=1.4"))
    .pipe(query, query=("Population <= 5000"))
    .pipe(query, query=("AveOccup <= 6"))
    .pipe(query, query=("MedHouseVal <= 5"))
    .assign(west_of_lon120=lambda d: d["Longitude"] <= -120)
    .assign(north_of_lat36=lambda d: d["Latitude"] >= 36)
)
```

```
start_pipeline | rows:   20,640 | cols:    9 | t:   0.0s | args:
query          | rows:   20,332 | cols:    9 | t:   0.0s | args: query: MedInc <= 10
query          | rows:   20,116 | cols:    9 | t:   0.0s | args: query: AveRooms <=10
query          | rows:   19,619 | cols:    9 | t:   0.0s | args: query: 0.8 <= AveBedrms <=1.4
query          | rows:   19,317 | cols:    9 | t:   0.0s | args: query: Population <= 5000
query          | rows:   19,232 | cols:    9 | t:   0.0s | args: query: AveOccup <= 6
query          | rows:   18,562 | cols:    9 | t:   0.0s | args: query: MedHouseVal <= 5
```



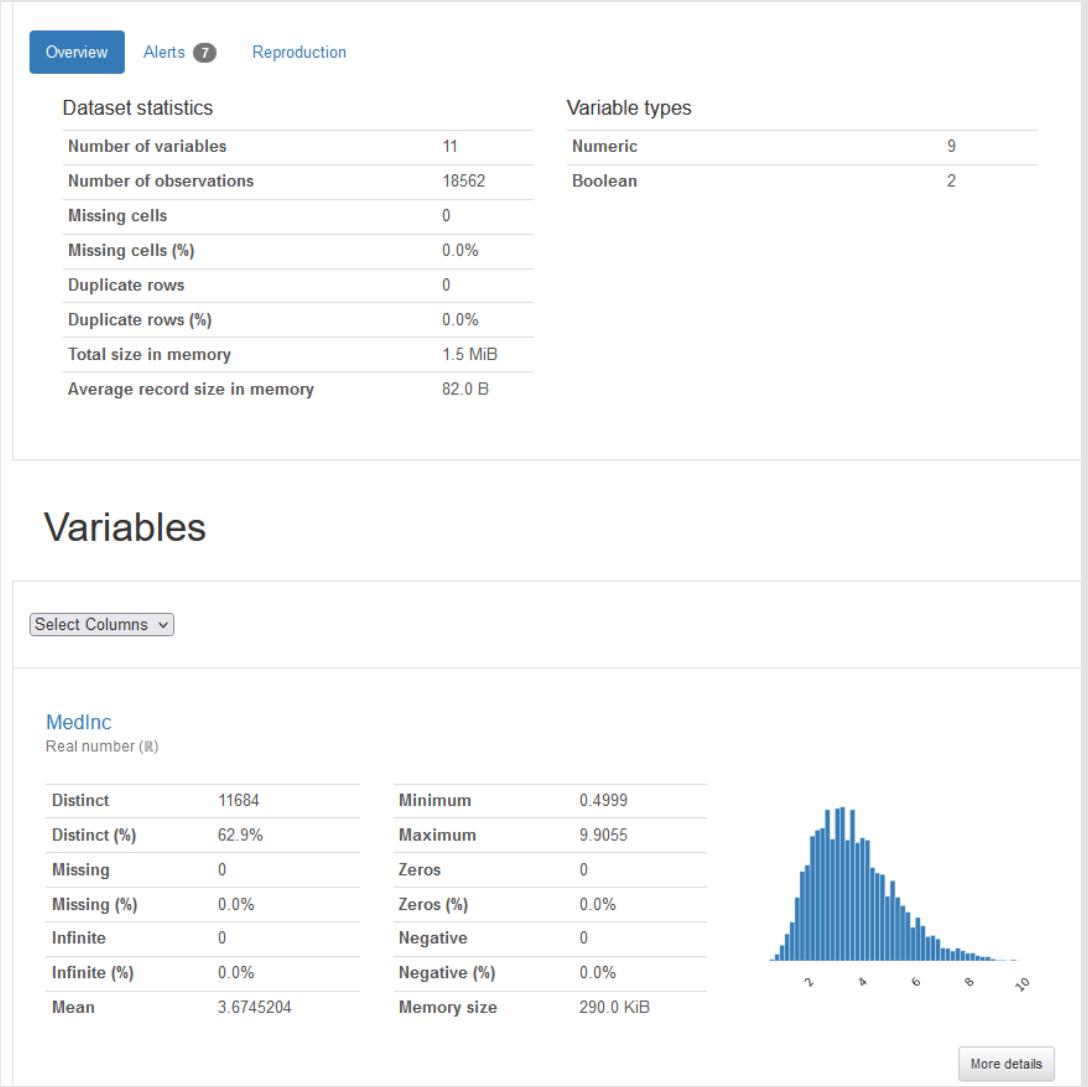
Automate Your EDA

ydata-profiling

github.com/ydataai/ydata-profiling

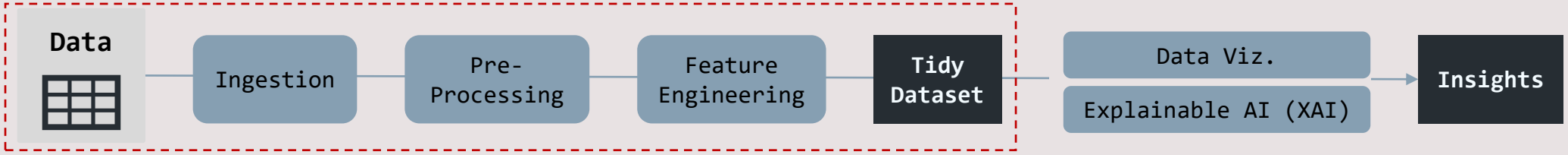
```
from ydata_profiling import ProfileReport

profile = ProfileReport(df)
profile.to_file("your_report.html")
```



Alternatives

1. **DataPrep:** github.com/sfu-db/dataprep#eda
2. **Sweetviz:** github.com/fbdesignpro/sweetviz



Explore Your Data Across Many Dimensions

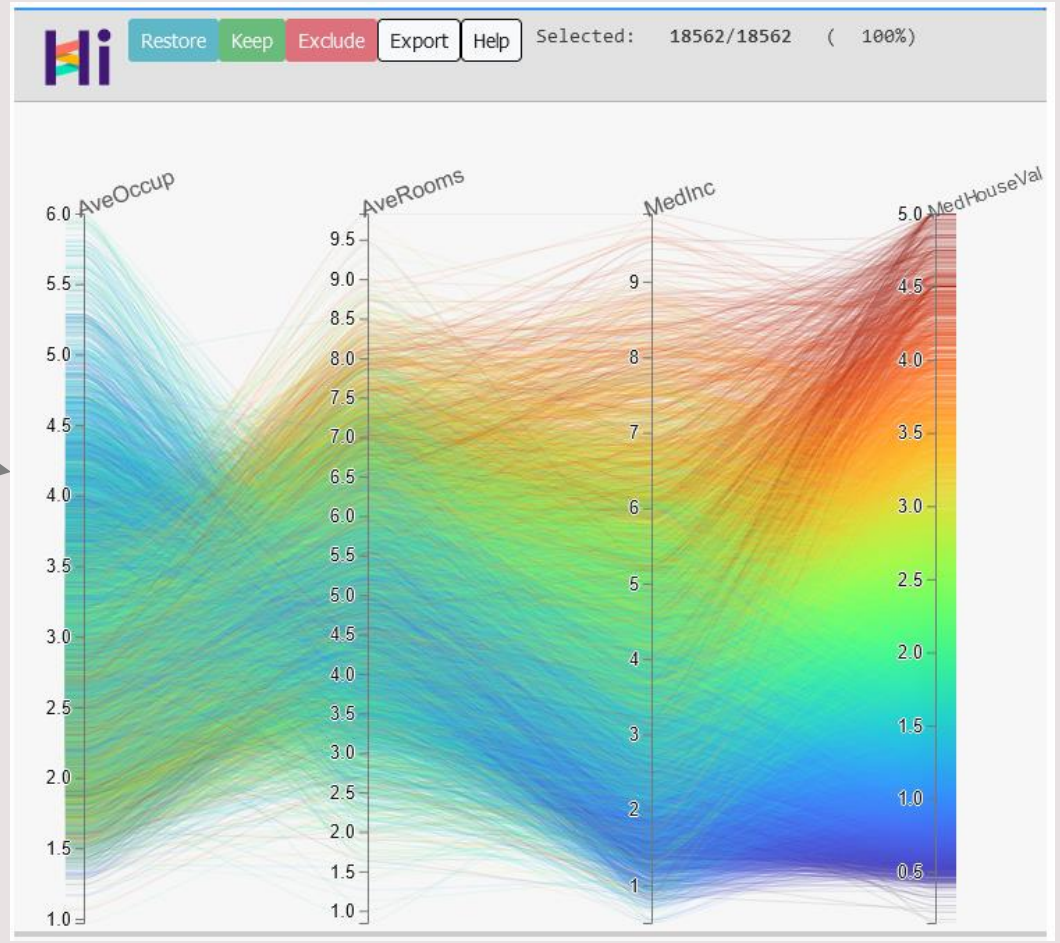
HiPlot

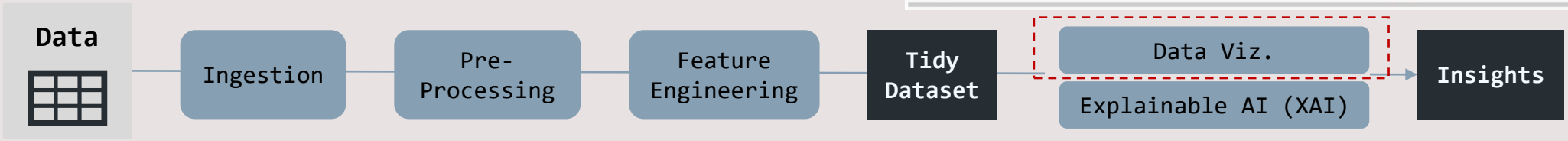
HiPlot

facebookresearch.github.io/hiplot

```
import hiplot

hiplot.Experiment.from_dataframe(df).display()
```





Highlight Patterns With Data Visualisation

PyGWalker

github.com/Kanaries/pygwalker

PyGWalker

```
import pygwalker

pygwalker.walk(df, axis=1))
```

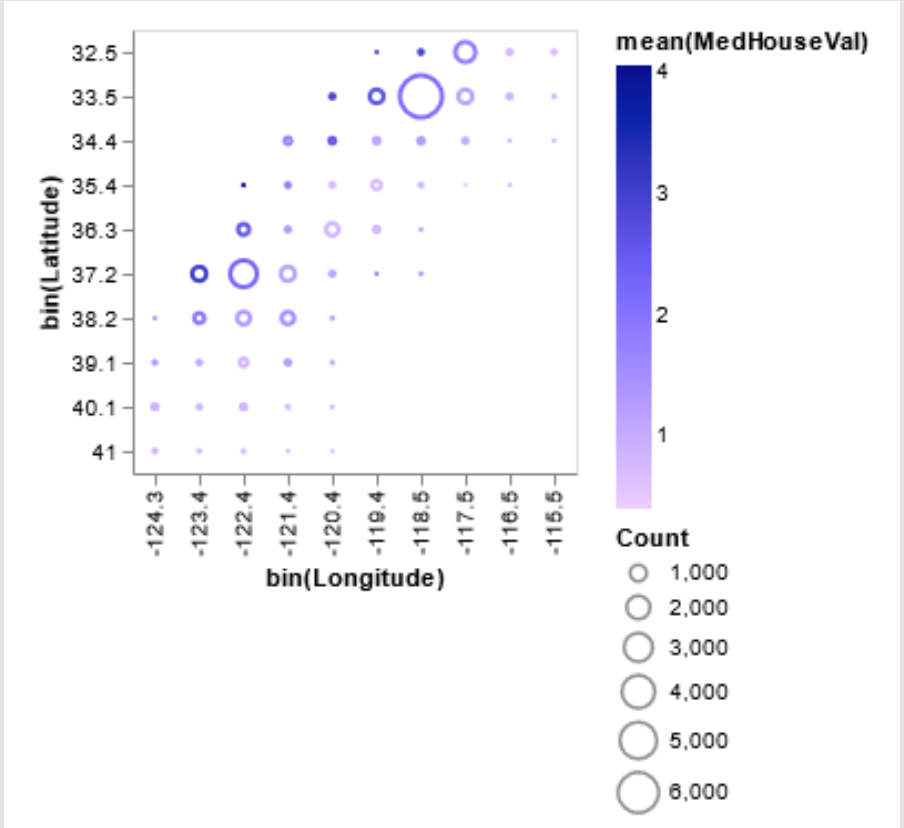
seaborn

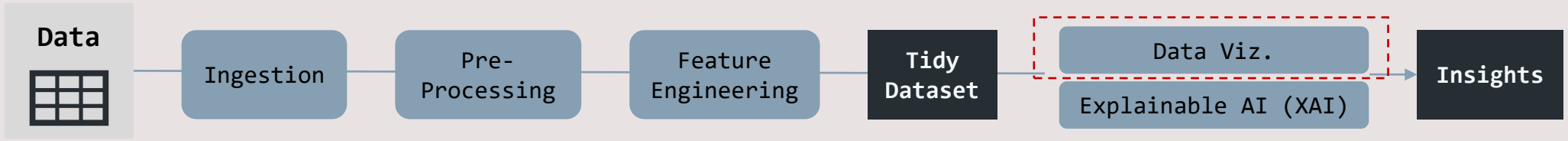
seaborn.pydata.org/tutorial.html

seaborn

```
import seaborn as sns

sns.relplot(data=df, ...)
sns.displot(data=df, ...)
sns.catplot(data=df, ...)
```





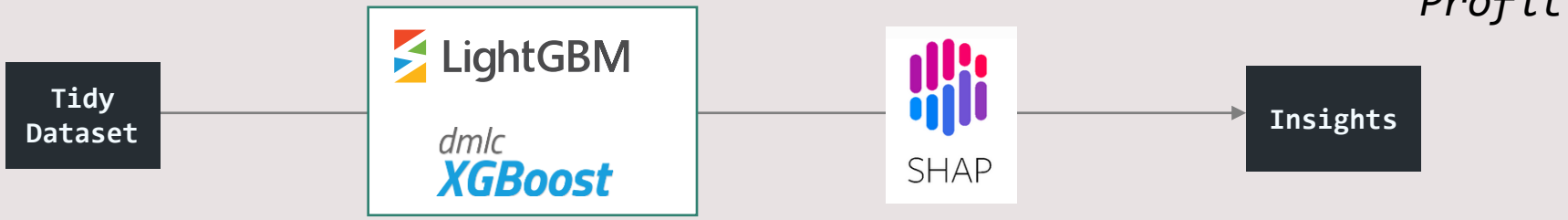
Insights from Explainable AI

I

Capture non-linear patterns with Gradient Boosted Decision Trees

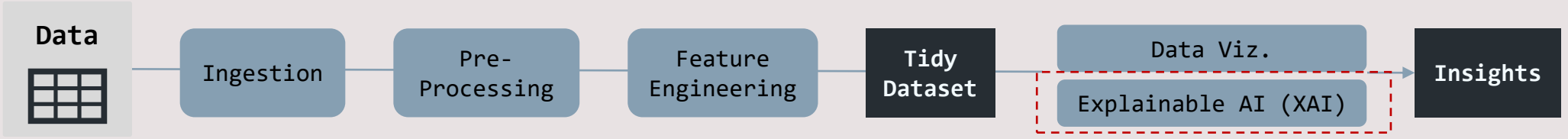
III

Profit



II

Highlight the patterns with SHAP

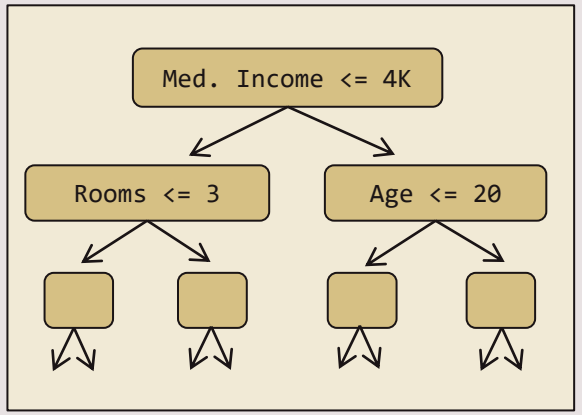


Capture Nonlinear Patterns With Gradient Boosted Decision Trees

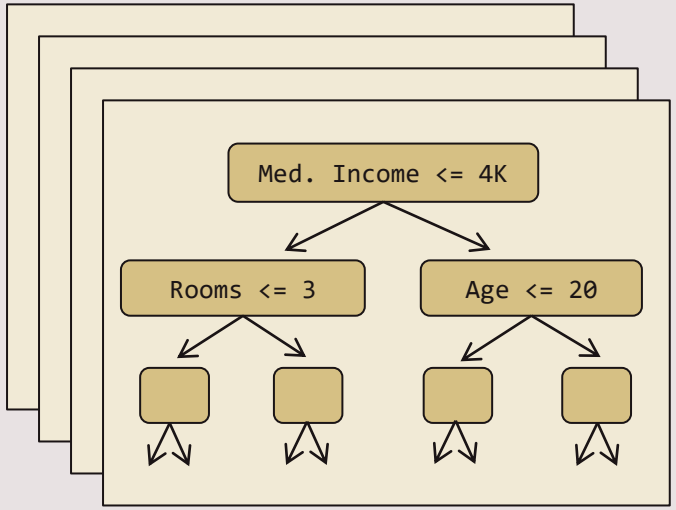
LightGBM

dmlc XGBoost

Decision Tree



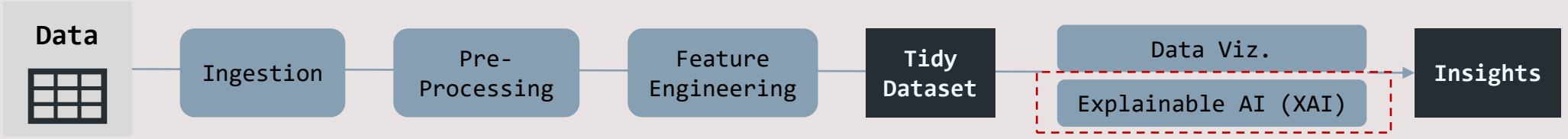
Gradient Boosted Decision Trees



\+ Best algorithm for most tabular datasets (benchmarks & Kaggle)
\+ Captures nonlinear patterns
\+ Great black-box predictors

\- Difficult to explain

More info: Pedro Tabacof - Unlocking the Power of Gradient-Boosted Trees (using LightGBM) | PyData London 2022



Capture Nonlinear Patterns With Gradient Boosted Decision Trees

Popular open-source implementations:

LGBM
LightGBM
lightgbm.readthedocs.io/en/latest

XGBoost
eXtreme Gradient Boosting
xgboost.readthedocs.io/en/stable

More info: Pedro Tabacof - Unlocking the Power of Gradient-Boosted Trees (using LightGBM) | PyData London 2022

Tracking experiments with MLflow

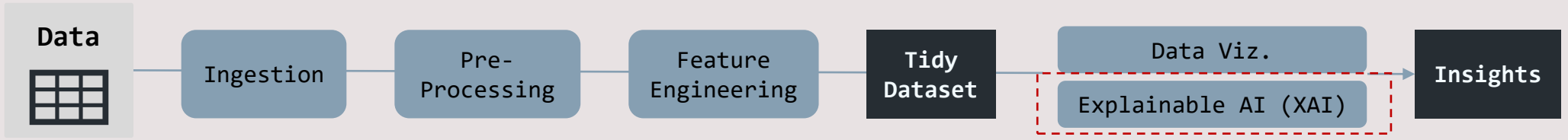
```
import mlflow

mlflow.autolog()
```

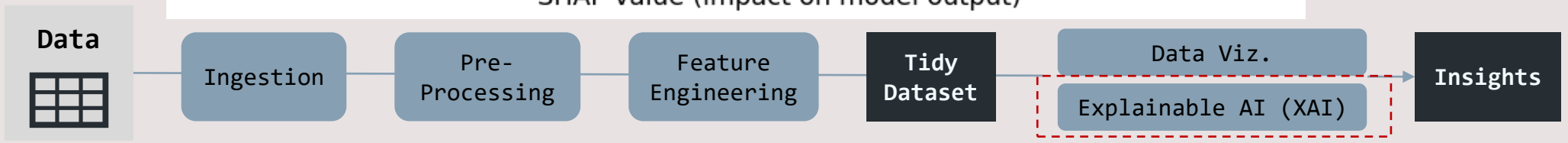
- Scikit-learn
- **XGBoost**
- **LightGBM**
- Keras
- Gluon
- Statsmodels
- Spark
- Fastai
- Pytorch

```
import mlflow

with mlflow.start_run():
    mlflow.log_param("x", 1)
    mlflow.log_metric("y", 2)
    ...
```

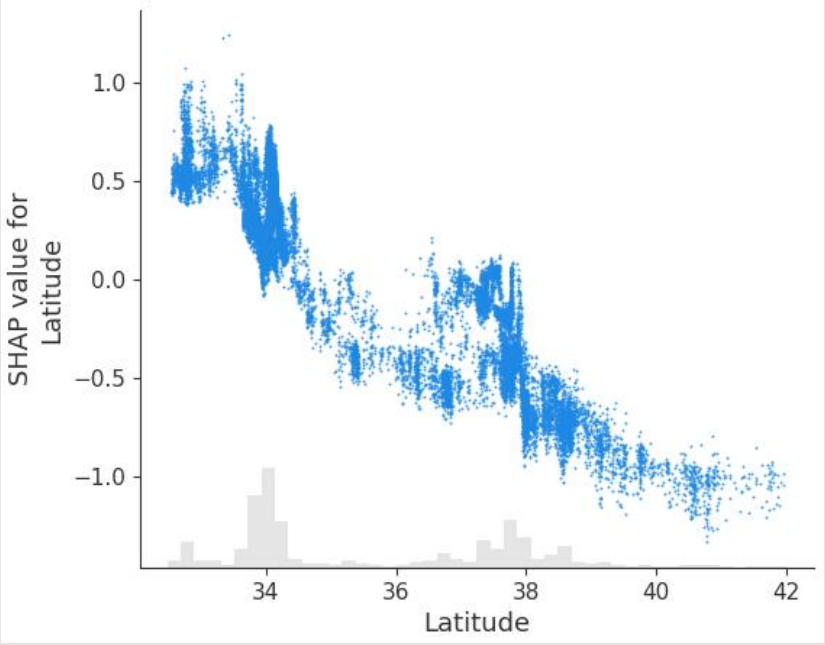


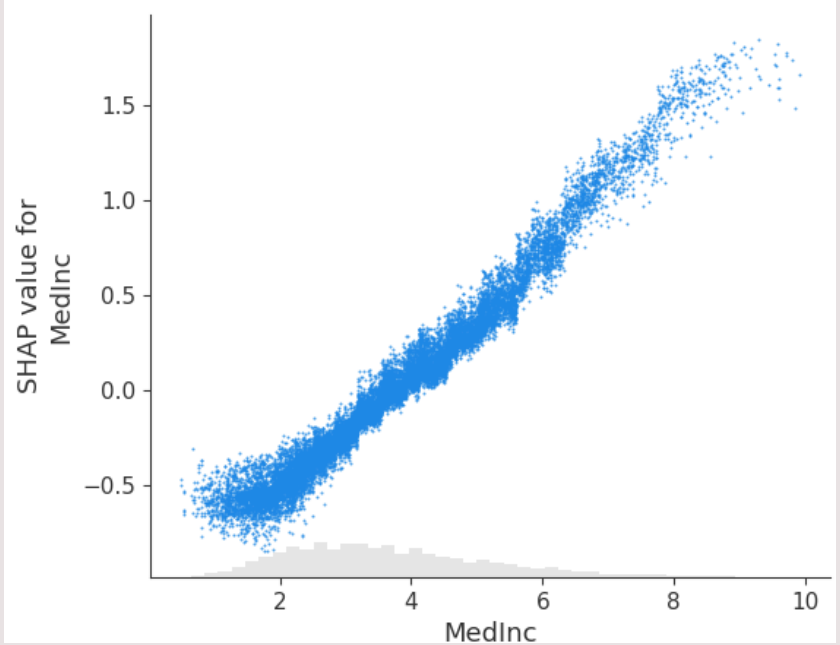
Insights from XAI: SHAP Summary Plot

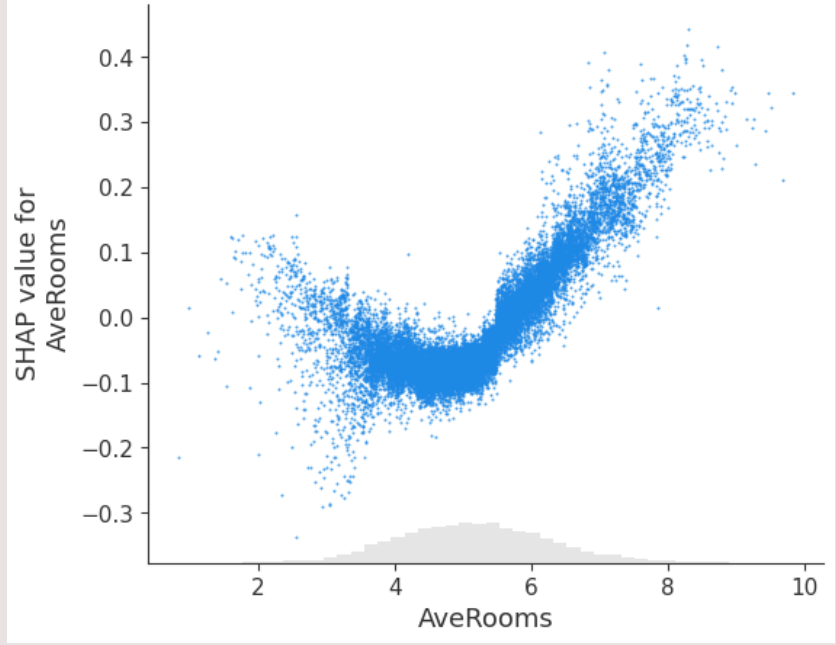


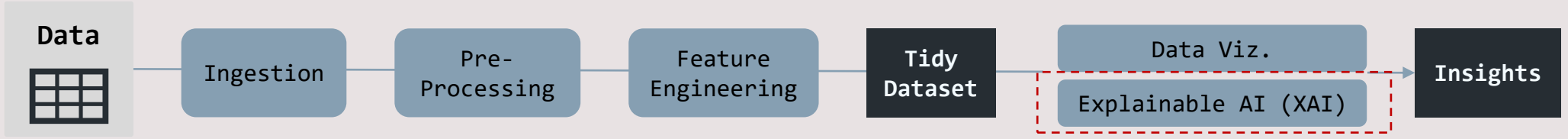
Source: https://shap.readthedocs.io/en/latest/index.html

Insights from XAI: SHAP Dependence Plots

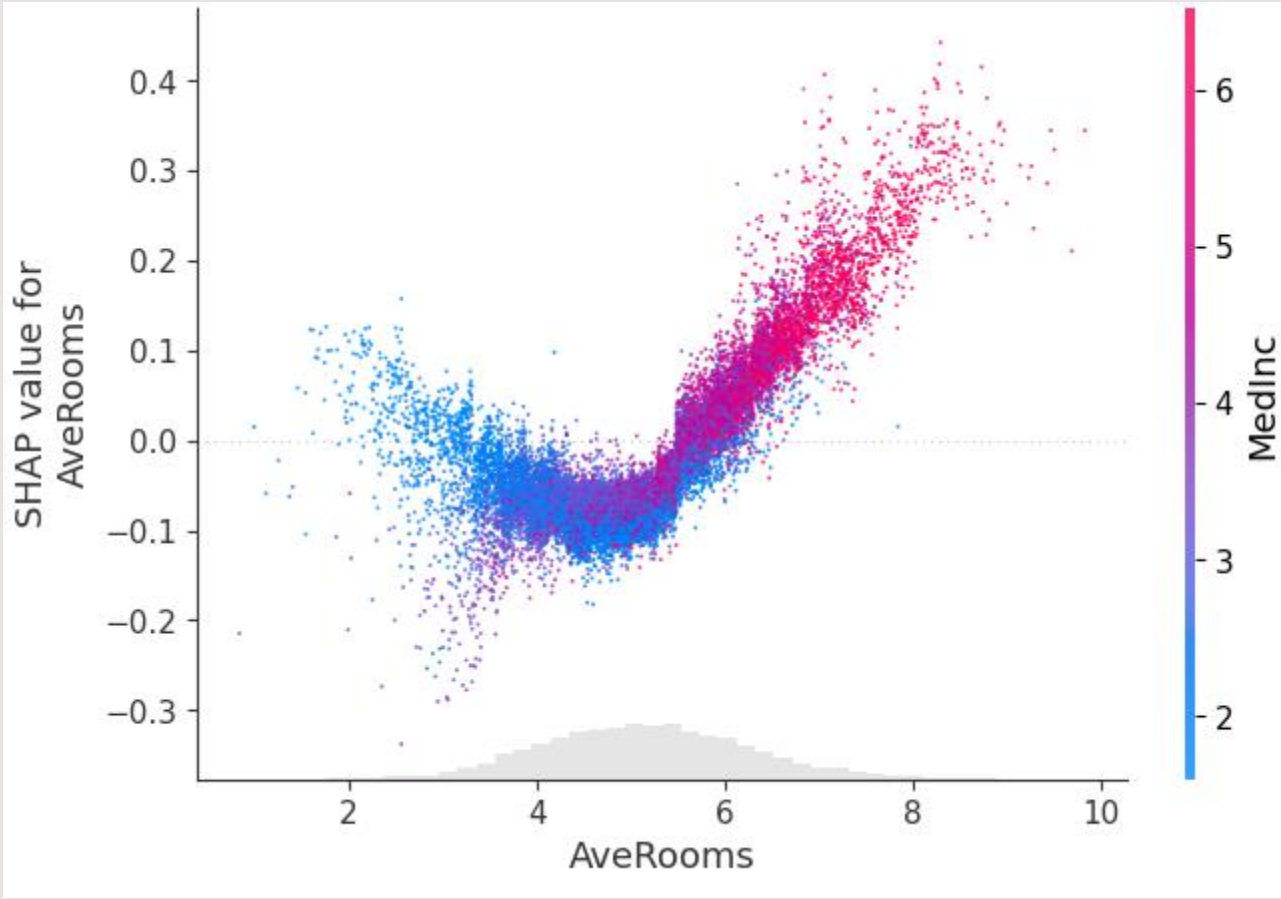


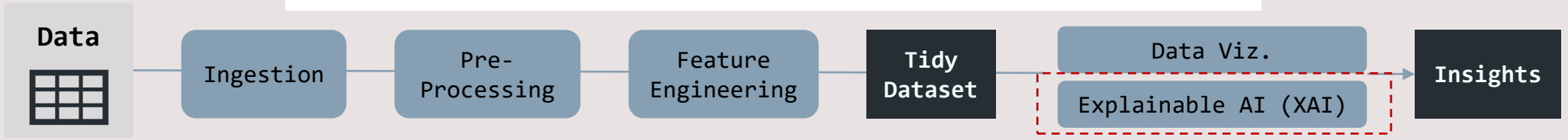




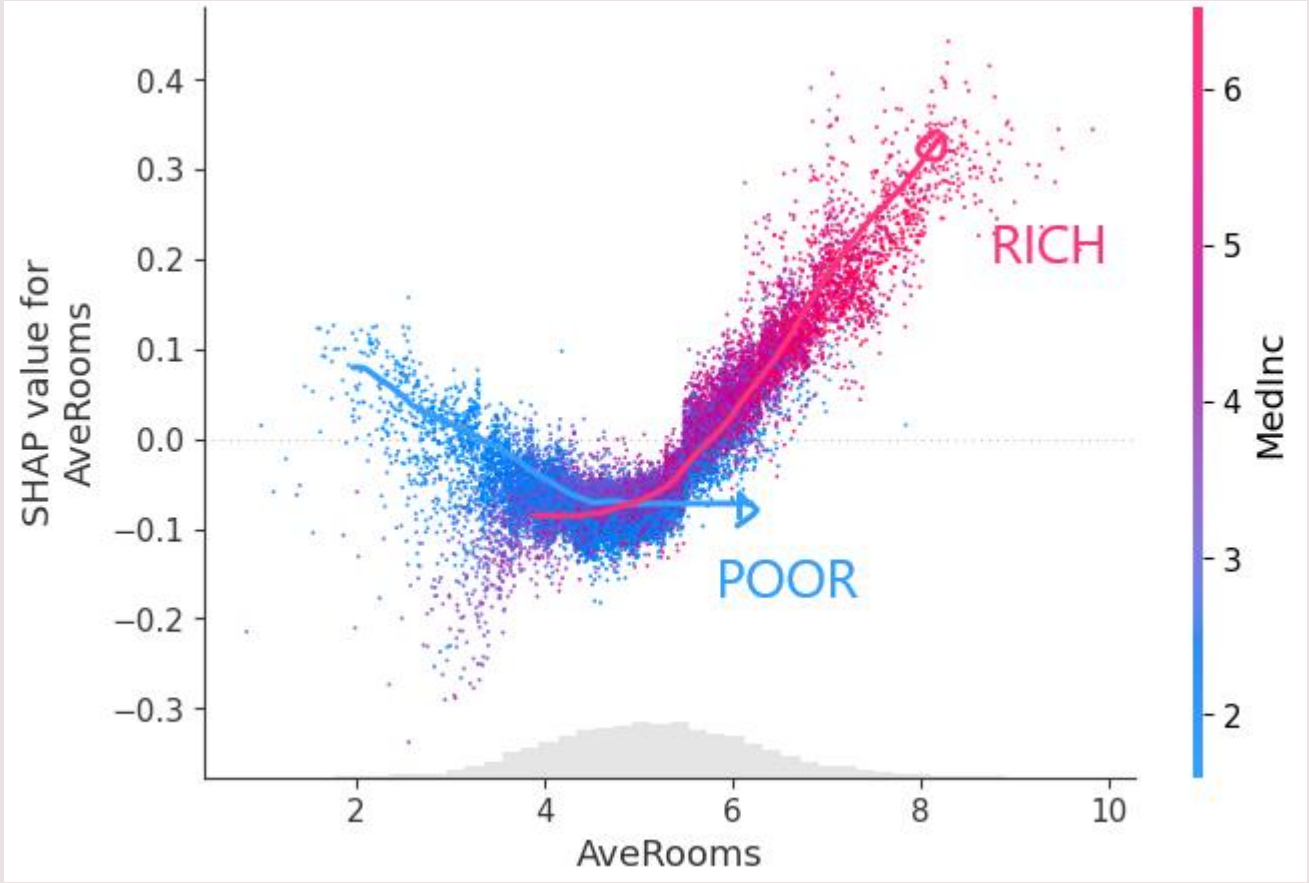


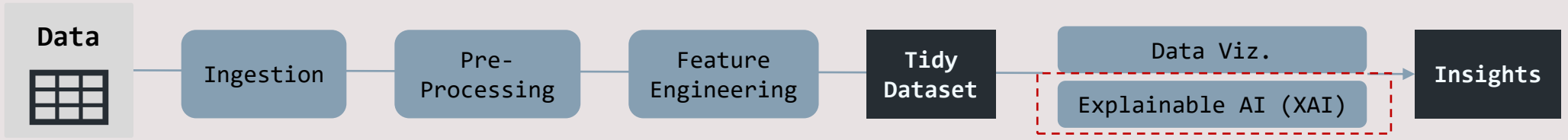
Insights from XAI: SHAP Dependence Plots





Insights from XAI: SHAP Dependence Plots

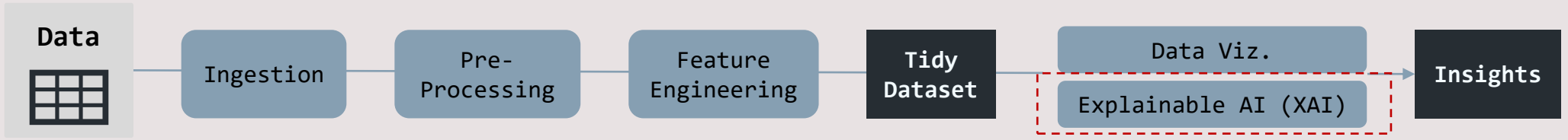




Managing the Complexity & Extracting Business Relevant Insights

Brainstorming session :

Raw Insights	Resulting Story	TODO: Next Iteration
1. Geography is important 2. Median income interacts with house size	A zone's median income changes how other variables impact house prices.	1. X 2. Y 3. Z



Agenda

1. Into : How to Create Value as a Data Scientist?
2. The SHAP Explainable AI Method
3. Demo: My Workflow on the California Housing Dataset
4. **Successful Applications in a Business Setting**

Successful Applications of XAI as a Product

Success stories:

1. Which populations were receptive to our marketing campaign ?
2. Which populations churn ? And what can we do about them ?
3. What are the different user patters on our app ?
4. Why is this ML model performing suspiciously well ?

Common business applications:

1. Diving into a new dataset and find non-linear patterns
2. Segmenting a population by some outcome, we wish to understand or influence
3. Reduce the uncertainty of a machine-learning projects

Data to Insights

Let’s Conclude...

What Should You Take Away From This Talk ?

These **mature open-source tools** will boost your productivity working with **tabular data**:

- panadas pipe
- Ydata profiling
- HiPlot / PyGWalker / Seaborn
- XGBoost / LightGBM
- MLFlow
- SHAP

You can extract insights by using **SHAP** (XAI) **on Gradient Boosted Decision Trees, resulting in:**

1. Great model performance
2. Transparency in what non-linear patterns are being picking up by the model in the data

Thank You for Your Attention!

Raphaël Lüthi

Machine Learning & AI Lead

@ Groupe Mutuel

Get in touch!

LinkedIn

Great resources to stay up to date:

- Ian Ozsvald's newsletter:
  - ✓ notanumber.email
- Jesper Dramsch's newsletter:
  - ✓ buttondown.email/jesper
- Vincent Warmerdam's tutorials:
  - ✓ calmcode.io
- Laszlo Sragner's code quality for data science community:
  - ✓ laszlo.substack.com
  - ✓ youtube.com/watch?v=FL6-XlRP7ZE
  - ✓ cq4ds.com
- Avi Chawla's daily blog:
  - ✓ blog.dailydoseofds.com

Thank You for Your Attention!