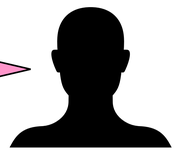
```
def combine(patients, ...):
 ...

def create_neural_net():
 ...

def execute_pipeline():
 ...
```

Original ML pipeline code

1 The data scientist provides the original pipeline code and specifies techniques to apply



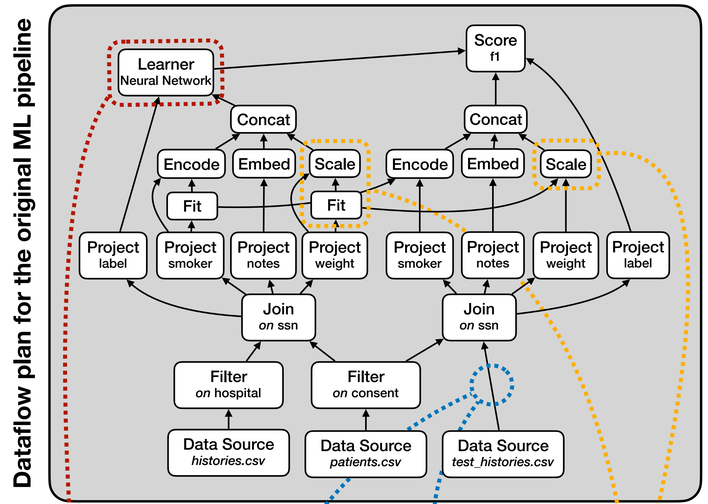
Data Scientist

```
analyses = [
 Cleanlearn(...),
 OperatorImpact(...),
]
```

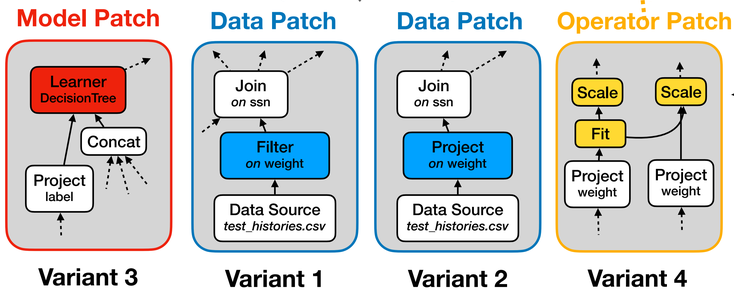
Declaratively specified data-centric analysis techniques

2 AutoDC extracts a dataflow plan from the original Python code via instrumentation

Plan Extraction via Instrumentation

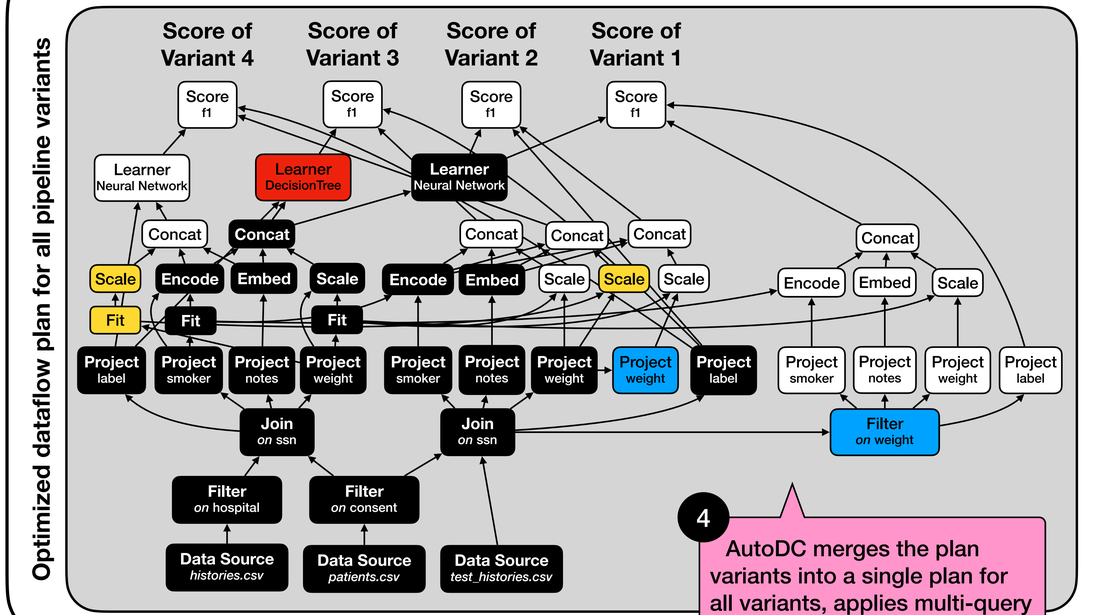


Variant Generation with Pipeline Patches



3 AutoDC generates plan variants for the analyses, based on pipeline patches

Multi-Query Optimization



4 AutoDC merges the plan variants into a single plan for all variants, applies multi-query optimizations and executes it

5 AutoDC generates a report per analysis for the data scientist

Report for operator impact analysis

operator change	f1_score
DecisionTree as model	80.3%
robust_scaler on weight	66.5%

Report for cleanlearn analysis

column	error	cleaning	f1_score
weight	outliers	filter	82.1%
weight	outliers	impute	69.7%