ORACLE

Machine Learning in Oracle Database

Eugene Simos, Witold Świerzy

Data Domain Experts

Oracle EMEA

September, 2022
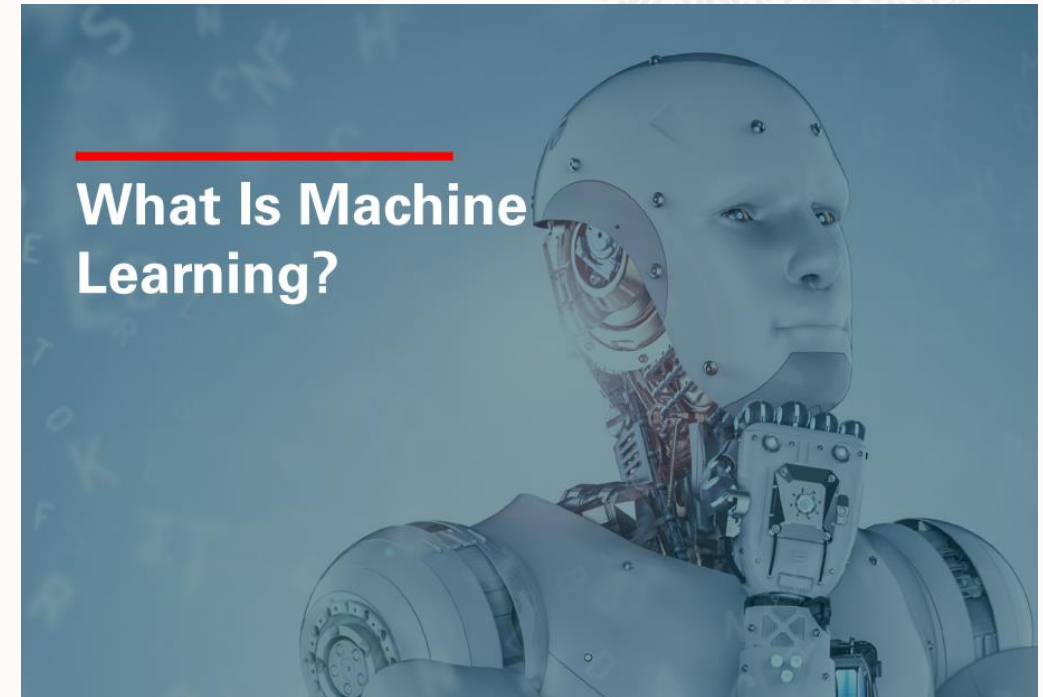
Machine Learning
Introduction

Wikipedia says:

- **Machine learning** (**ML**) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.

- It is seen as a part of **Artificial Intelligence**.

- **Machine learning algorithms** build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

What Is Machine Learning?

Machine Learning Use Cases

Industry

Banking

- Risk Management (Fraud/Credit Risk)
- Digital Banking Experience
- Customer Loyalty and Retention
- Macro/Micro Segmentation
- Next Best Offer
- Cross & Upsell
- Profitability and Revenue Forecasting

Telco & Media

- AI Driven Network Operations
- Marketing and Sales Automation
- Predictive Maintenance
- Offer/Package Personalization and Recommendation
- Customer Value and Experience (Churn, Retention, Loyalty, Segmentation, etc.)
- Smart Media Production Planning
- Data monetization

Healthcare & Public Sector

- Population Health
- Chronic Disease Management
- Clinical Decision Support
- Fraud Prevention
- Payment and Claim Management
- Citizen Analytics
- Improved Public Services/Healthcare-Education-Finance-Transportation-Social Intelligence
- Crime Cyber Thread Prediction

Retail

- Personalized Offerings
- Location Based Analytics
- Brand Sentiment Analysis
- Next Best Offer
- Customer Loyalty and Retention
- Segmentation
- Demand Prediction
- Order Picking and Logistic Optimization
- Inventory Management

Machine Learning Use Cases

Industry

Manufacturing
- Predictive & Condition Based Maintenance
- Demand Forecasting & Inventory Management
- Product Quality Analytics
- Risk Analytics and Safety Improvements
- Revenue Prediction
- Price Optimization
- Sustainability Optimization & Improvements

Oil & Gas
- Demand Forecasting
- Operational Risk/Environment
- Process Regulations
- Logistics and Operations
- Energy Trading
- Cash flow Mgmt.
- Credit Risk Mgmt.

Insurance
- Customer Segmentation
- Cross Sell/Upsell
- Campaign Optimization
- Regulatory Reporting
- Claims Analytics
- Telematics
- Risk Analysis
- Catastrophe Modeling
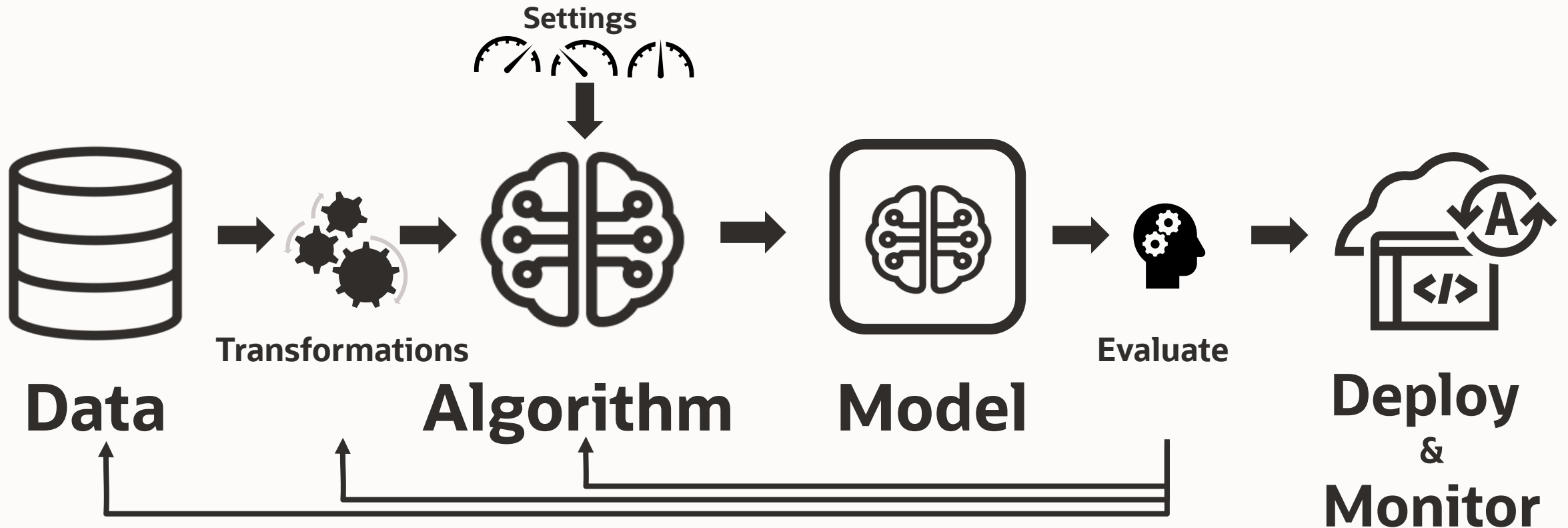- Claims Fraud
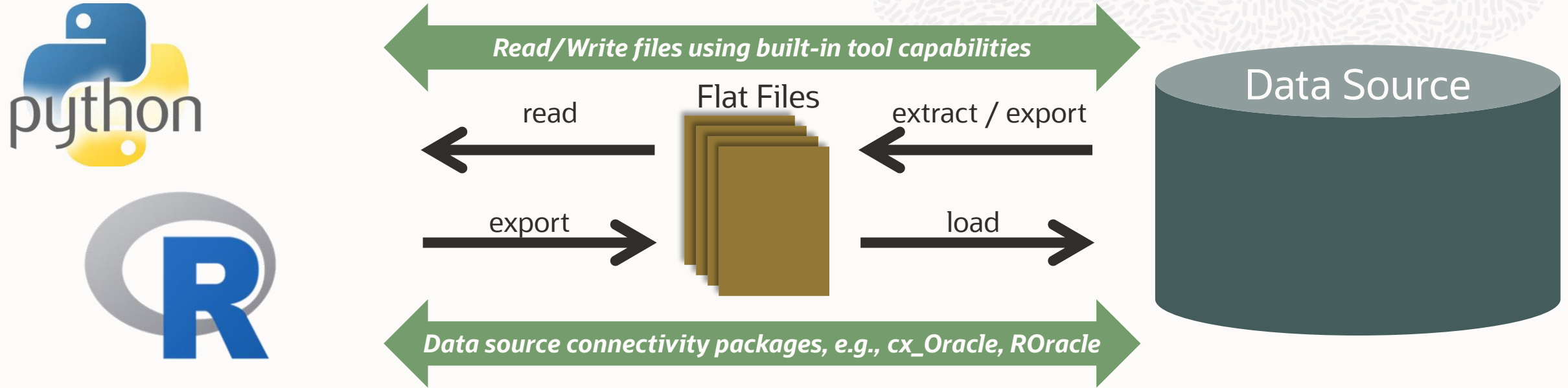- Underwriting Fraud
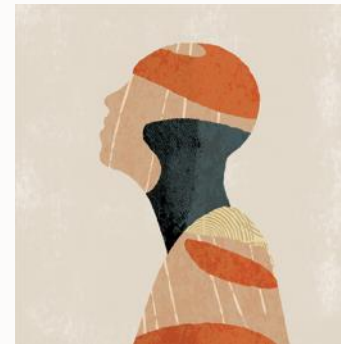
Automobile
- Predictive Maintenance
- Customer 360
- Dealer Operations and Parts Optimization
- Shared Mobility Services
- Multimodal Transportation
- Warranty Claims
- Loan Monitoring
- Collections
- Consumer Credit Scoring

Traditional approaches to ML modeling and deployment are cumbersome

Settings

Data

Transformations

Algorithm

Model

Evaluate

Deploy & Monitor

Traditional Python and Data Source Interaction

Read/Write files using built-in tool capabilities

read

Flat Files

extract / export

export

load

Data Source

Data source connectivity packages, e.g., cx_Oracle, ROracle

Access latency
Paradigm shift: Python → *Data Access Language* → Python/R
Memory limitation – data size, in-memory processing
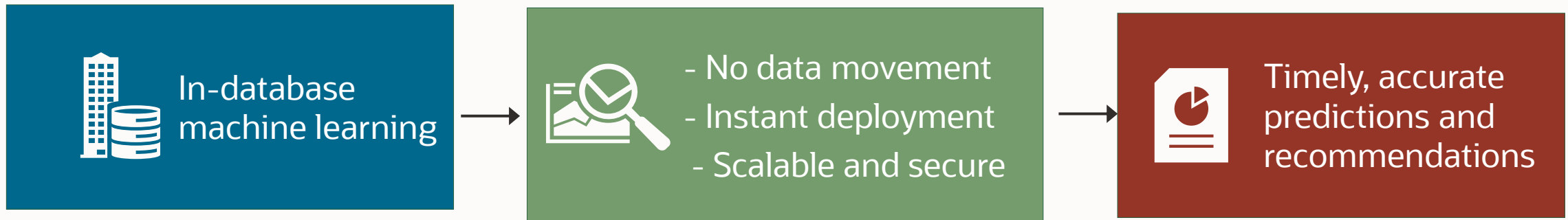Single threaded
Issues for backup, recovery, security
Ad hoc production deployment

What if Machine Learning could be easier?

In-database machine learning makes ML
modeling and deployment faster, more secure

| In-database machine learning | - No data movement
- Instant deployment
- Scalable and secure | Timely, accurate predictions and recommendations |

Stay one step ahead of the competition

Oracle Machine Learning

Automated

Automated machine learning supports data scientist productivity and empowers non-experts

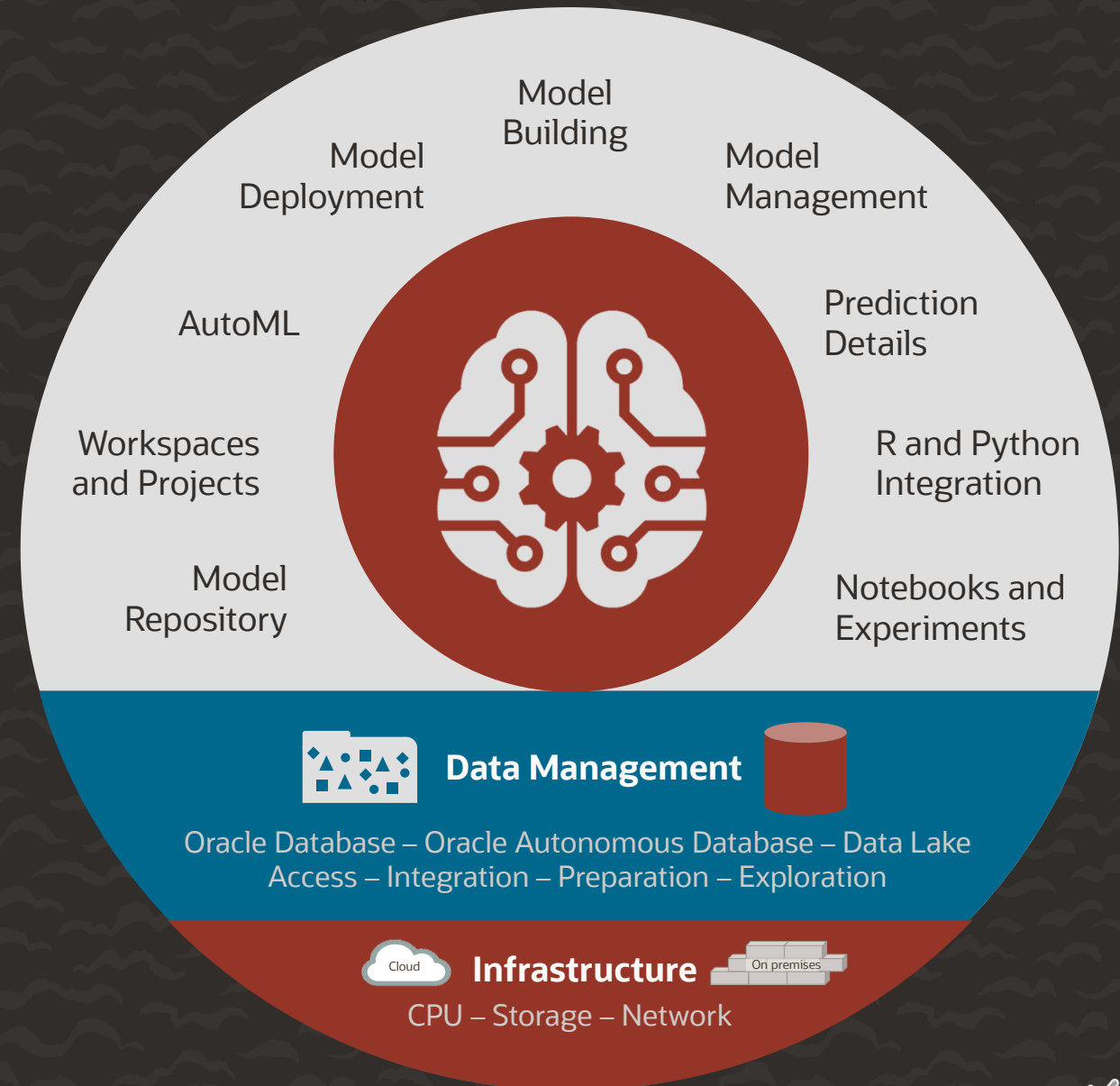Algorithm-specific data preparation, integrated text mining, partitioned models

Scalable

Over 30 high performance, parallelized in-database machine learning algorithms that require no data movement

Production-ready

Quickly deploy and update machine learning models in production via SQL and REST APIs

Real-time predictions for end-user applications and dashboards

Deploy R and Python user-defined functions using managed processes with easy data-parallel and task-parallel invocation

Model Building

Model Deployment

Model Management

AutoML

Prediction Details

Workspaces and Projects

R and Python Integration

Model Repository

Notebooks and Experiments

Data Management

Oracle Database – Oracle Autonomous Database – Data Lake
Access – Integration – Preparation – Exploration

Cloud **Infrastructure** On premises

CPU – Storage – Network

Oracle Machine Learning on RAC

Scalability and Performance

- **Familiar algorithms redesigned to enable distributed parallelism and scalability across cluster nodes**

- **Scoring takes advantage of storage-tier optimizations with function push-down (Exadata platform)**

- **Optimized memory utilization**

  - Data brought into memory incrementally as needed

  - Models cached in the library cache and can be shared across queries

  - Leverages disk-aware structures – relying on DB memory manager for efficient allocation in multi-user environment

  - When building/scoring partitioned models, not all partitions need be loaded

Oracle Machine Learning

OML4SQL   **OML Notebooks**

Collaborative notebook environment based on Apache Zeppelin with Autonomous Database

OML4R   **Oracle Data Miner**

Interfaces for 3 popular data science languages: SQL, R, and Python

SQL Developer extension to create, schedule, and deploy ML solutions through a drag-and-drop interface

OML4Py   **OML4Spark**

No-code AutoML interface on Autonomous Database

OML AutoML UI

ML for the big data environment from R with scalable algorithms

OML Services

Model Deployment and Management, Cognitive Text on Autonomous Database

Supervised Learning vs Unsupervised Learning
Two categories of Machine Learning algorithms

Supervised Learning

Has a known answer (target) to learn from and compare against

- **Classification** predicts a discreet target
- **Regression** predicts a continuous numeric target
- **Time Series Forecasting** predicts time-dependent future values based on previously observed values

Terms related to this category

Target is what you're trying to predict

Predictors/variables/features are data used to predict the target

Case ID is the unique identifier per row/example/case

ML Model is the representation of the patterns found in the data, that can be used to serve a particular purpose in the context of an algorithm

[OML Algorithm Cheat Sheet](#)

Unsupervised Learning

Finds patterns, but no known answer to learn from or compare against

- **Clustering** groups objects such that objects in the same group are more similar to each other, than to those in other groups (clusters)
- **Market Basket Analyzis** (Associacion Rules) discovers the rules, that determine how or why certain items are connected: happen, used or purchased together
- **Anomaly Detection** finds usual cases in data, that may warrant further exploration or investigation
- **Feature Extraction** creates derived values (feature) used in subsequent ML steps , may concentrate the " signal" in the data

Oracle Machine Learning Algorithms and Analytics in Oracle Database

CLASSIFICATION
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine (SVM)
- Explicit Semantic Analysis
- *XGBoost**

ANOMALY DETECTION
- One-Class SVM
- *MSET-SPRT**

CLUSTERING
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

TIME SERIES
- Forecasting - Exponential Smoothing
- Includes popular models
  e.g. Holt-Winters with trends,
  seasonality, irregular time series

[OML Algorithm Cheat Sheet](#)

REGRESSION
- Generalized Linear Model (GLM)
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- *XGBoost**

ATTRIBUTE IMPORTANCE
- Minimum Description Length
- Random Forest
- Unsupervised Pairwise KL Divergence
- CUR decomposition for row & AI

ASSOCIATION RULES
- A priori

PREDICTIVE QUERIES
- Predict, cluster, detect, features

SQL ANALYTICS
- SQL Windows
- SQL Patterns
- SQL Aggregates

FEATURE EXTRACTION
- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

ROW IMPORTANCE
- CUR Decomposition

RANKING
- *XGBoost**

TEXT MINING SUPPORT
- Algorithms support text columns
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA)

STATISTICAL FUNCTIONS
- min, max, median, stdev, t-test, F-test,
  Pearson's, Chi-Sq, ANOVA, etc.

*Includes support for Partitioned Models,
Transactional data and aggregations*

* New in 21c

Oracle Machine Learning interfaces and platform availability

IDE	OML Component	Oracle Autonomous Database (19c, 21c)	Oracle Database (19c, 21c)	Oracle DBCS	Oracle Exadata CI/C@C
SQL Developer SQL*Plus	**OML4SQL**	✔ ADB-S, ADB-D, ADB C@C	✔	✔	✔
Python client PyCharm, Zeppelin, Jupyter	**OML4Py**	✔ ADB-S	✔	✔	✔
R client RStudio (Server)	**OML4R**	✔ ADB-S	✔	✔	✔
Zeppelin Notebook	**OML Notebooks** (SQL, PL/SQL, Python, markdown)	✔ ADB-S			
No-code UI	**OML AutoML UI**	✔ ADB-S			
REST API clients Postman	**OML Services**	✔ ADB-S			
No-code UI in SQL Developer	**Oracle Data Miner**	✔ ADB-S, ADB-D, ADB C@C	✔	✔	✔

Oracle Machine Learning for SQL

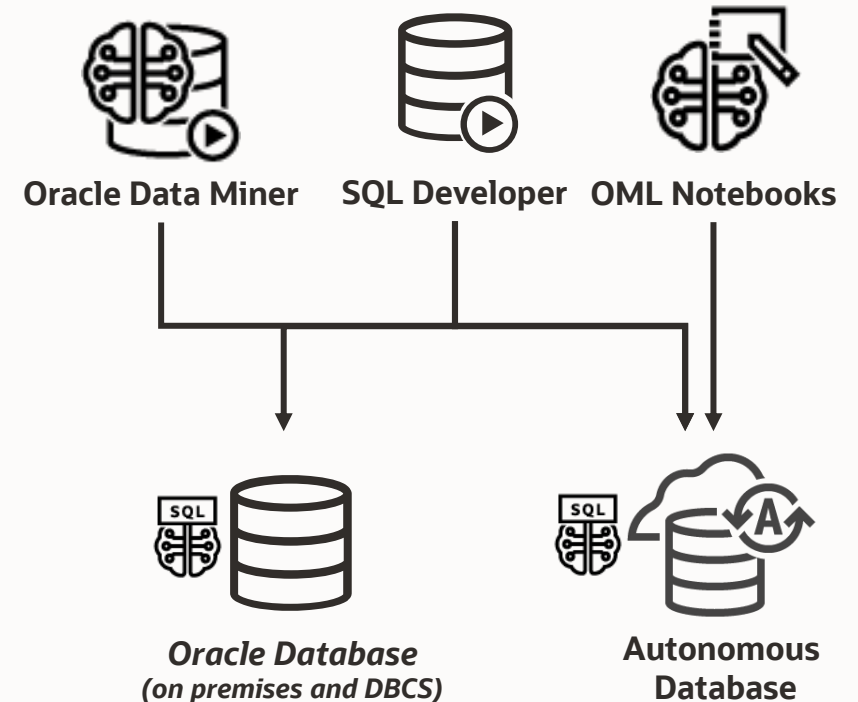Empower SQL users with ML included in Oracle Database

In-database, parallelized, distributed algorithms

- No extracting data to separate ML engine
- Fast and scalable
- Batch and real-time scoring at scale that leverages Exadata storage-tier function pushdown
- Algorithm-specific automatic data preparation
- Explanatory prediction details

ML models as first-class database objects

- Access control per model
- Audit user actions
- Export / import models across databases
- Ease of backup, recovery, and security

Faster time-to-market through immediate solution deployment

Oracle Data Miner   **SQL Developer**   **OML Notebooks**

Oracle Database
(on premises and DBCS)

Autonomous Database

Oracle Machine Learning for Python and R

Empower data scientists with open source environments

Oracle Database as HPC environment

Manipulating database data at scale for data preparation/exploration, machine learning modeling and production deployment

In-database parallelized and distributed machine learning algorithms
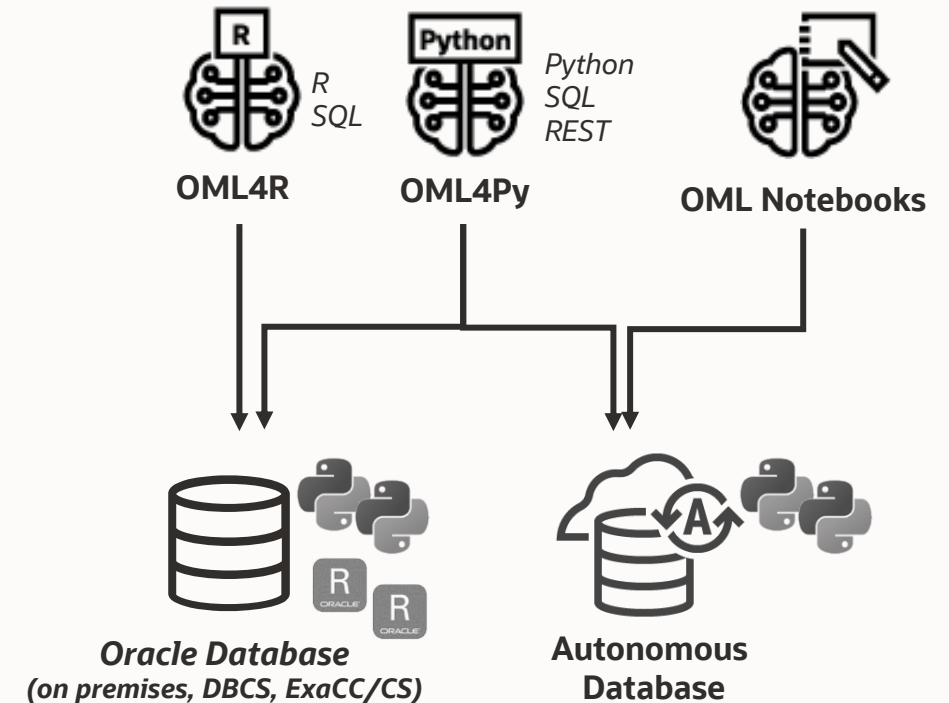
Manage scripts and objects in Oracle Database

Integrate results into applications and dashboards via SQL or REST

Embedded execution to run user-defined R and Python functions in engines spawned by the database environment

OML4Py automated machine learning

- Automated algorithm and feature selection
- Automated model tuning and selection

Oracle Machine Learning for Python and R

Empower data scientists with open source environments

Transparency layer
- Leverage proxy objects so data remains in database
- Overload native functions translating functionality to SQL
- Use familiar Python/R syntax on database data
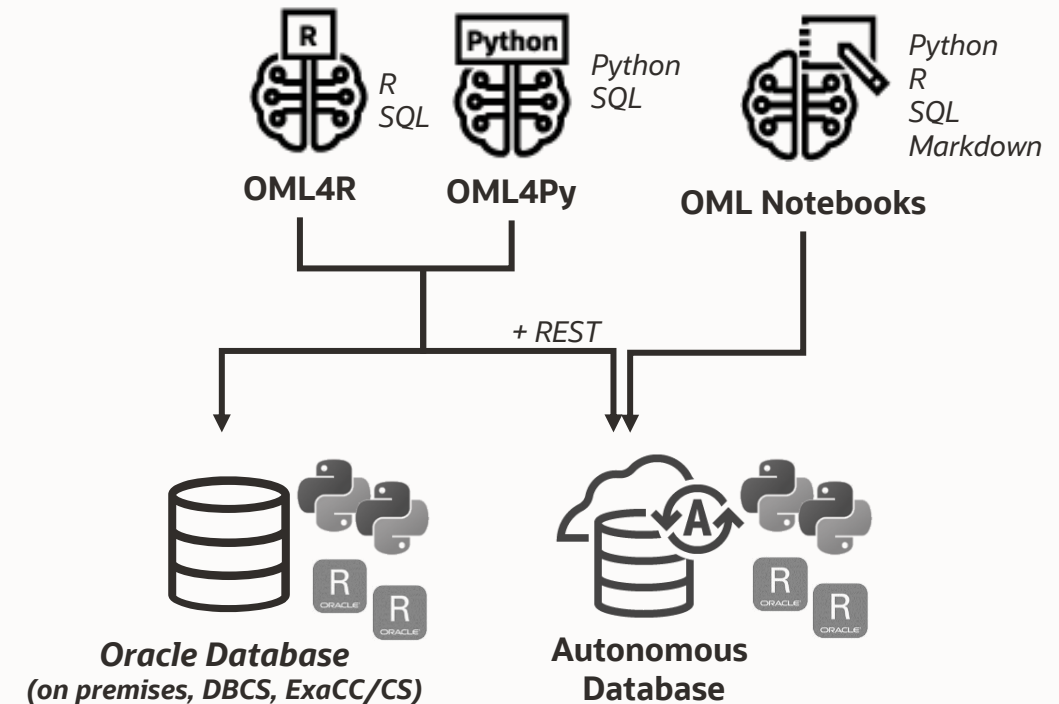
Parallel, distributed in-database algorithms
- Scalability and performance
- Exposes in-database algorithms available from OML4SQL

Embedded execution
- Manage and invoke Python or R scripts in Oracle Database
- Data-parallel, task-parallel, and non-parallel execution
- Use open source packages to augment functionality

OML4Py AutoML and MLX
- Automated algorithm selection, feature selection, model tuning
- Algorithm-agnostic model explainability (MLX) for feature ranking

OML4R   *R*
SQL

OML4Py   *Python*
SQL

OML Notebooks   *Python*
R
SQL
Markdown

+ REST

Oracle Database
(on premises, DBCS, ExaCC/CS)

Autonomous Database

Oracle Data Miner User Interface

Create analytical workflows – productivity tool for data scientists – enables citizen data scientists
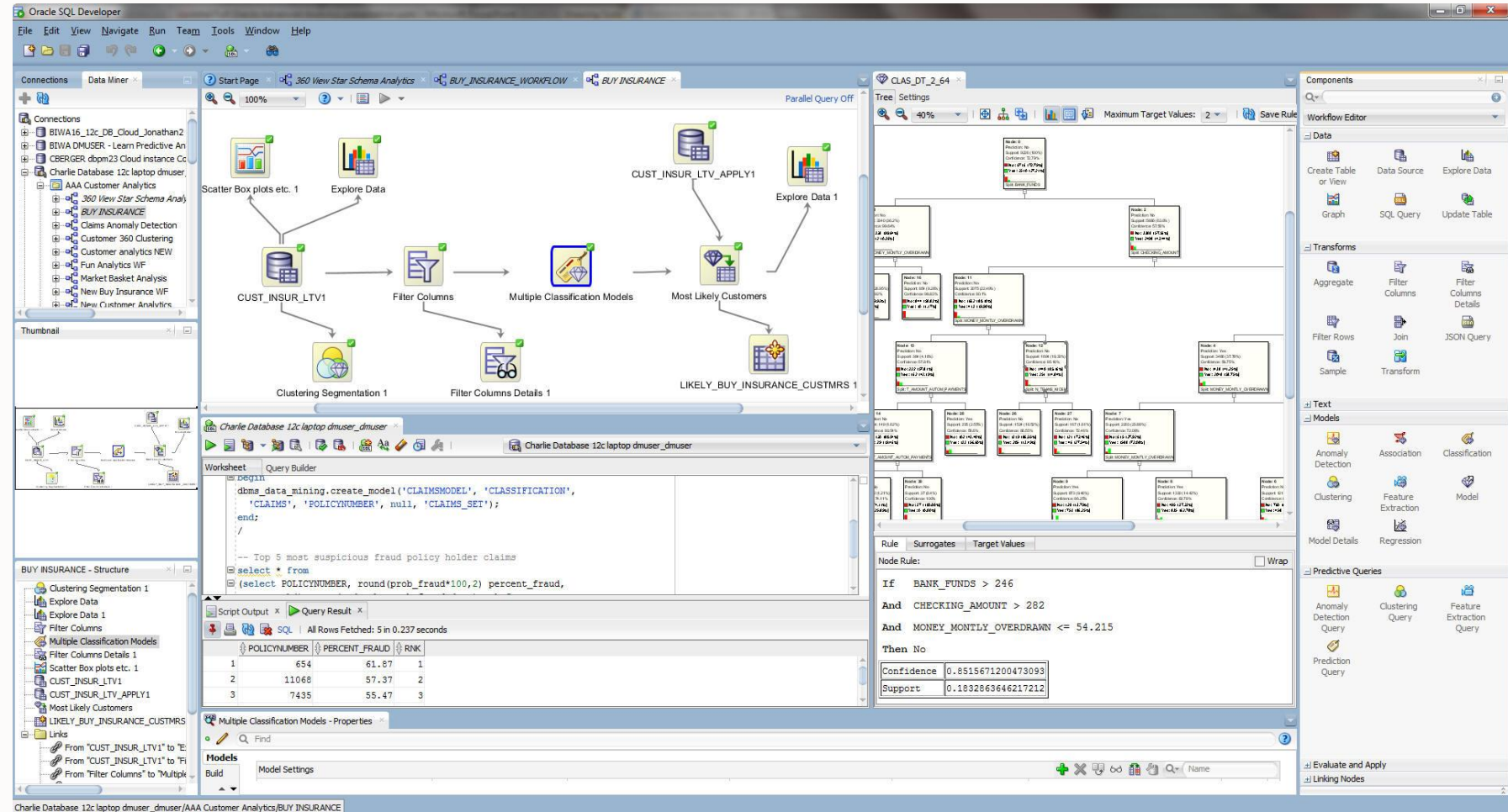
SQL Developer Extension

Automates typical data science steps – data exploration and preparation, ML modeling

Easy to use drag-and-drop interface

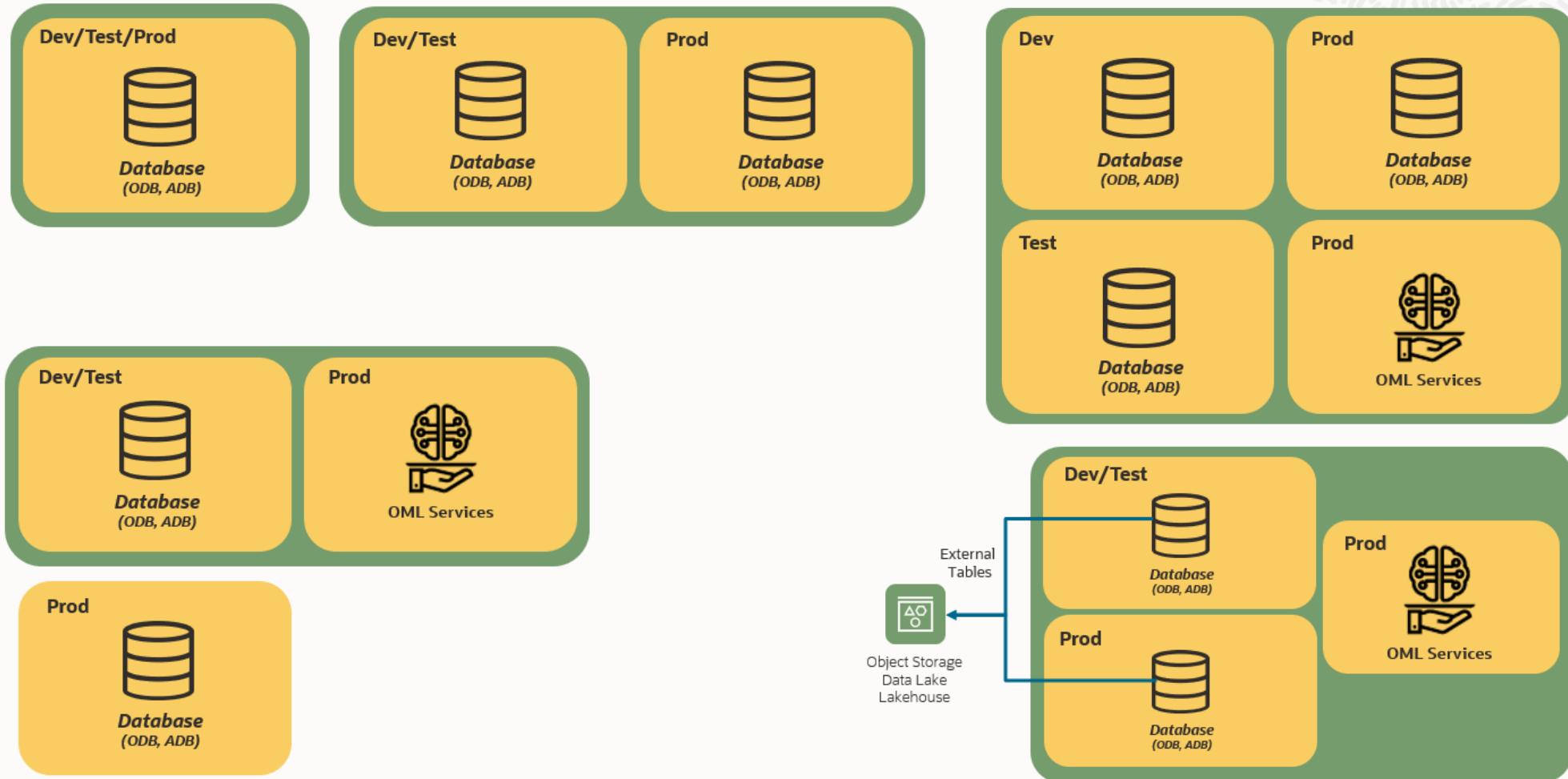Analytical workflows quickly defined and shared

Wide range of algorithms and data transformations

Generate SQL code for immediate deployment

Deployment Architectures
In-database models

Conversations with Customers

Questions to ask a prospective customer

- **The big picture**
  - Type of the business
  - Areas of improvements,
  - Main painpoints of current solution
- **Type of solution**
  - Packaged application or raw technology
  - Results of ML project consumption
- **Data**
  - What type of data the customer has
  - Volume of the data
  - Who manages the data
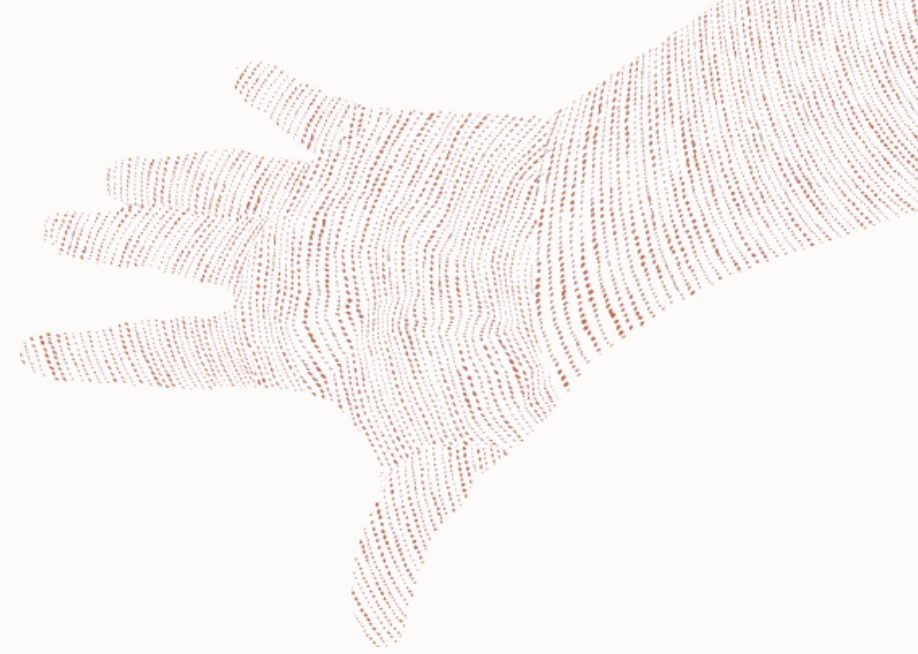 - Storage type

Summary
Key Points

During this section we

- Defined Machine Learning

- Discussed various use cases of ML

- Presented benefits of using Oracle ML over traditional approch

- Shortly reviewed list of ML algorithms available in Oracle Database

- Described all Oracle Machine Learning APIs available in Oracle Database

- Discussed various deployment architectures of Oracle Machine Learning

Questions?

Thank you