End-to-End Data Science using IBM's Cloud Pak for Data

January 14, 2021

The session starts at 9:00am.

End-to-End Data Science using IBM's Cloud Pak for Data

Power of data. Simplicity of design. Speed of innovation.

Bernie Beekman
Michael Cronk

Agenda

Time	Description
09:00 AM – 10:00 AM	**Overview of Cloud Pak for Data** **Lab Orientation 1,2**
10:00 AM – 11:45 AM	**Lab-1: Set up Environment, Lab-2: Watson Knowledge Catalog**
11:45 AM – 12:15 PM	**Lab Review 1,2 /Lab Orientation 3,4** **Lunch**
12:15 PM – 02:00 PM	**Lunch** **Lab-3: Data Refinery, Lab-4: SPSS Modeler**
02:00 PM – 02:30 PM	**Lab Review 3,4 / Lab Orientation 5,6**
02:30 PM – 03:30 PM	**Lab-5: Machine Learning with SparkML, Lab-6: AutoAI**
03:30 PM – 04:00 PM	**Lab Review 5,6 / Lab Orientation 7,8**
04:00 PM – 05:00 PM	**Lab-7 – Watson OpenScale, Lab-8 Decision Optimization**
05:00 PM – 05:15 PM	**Lab-Review 7,8**

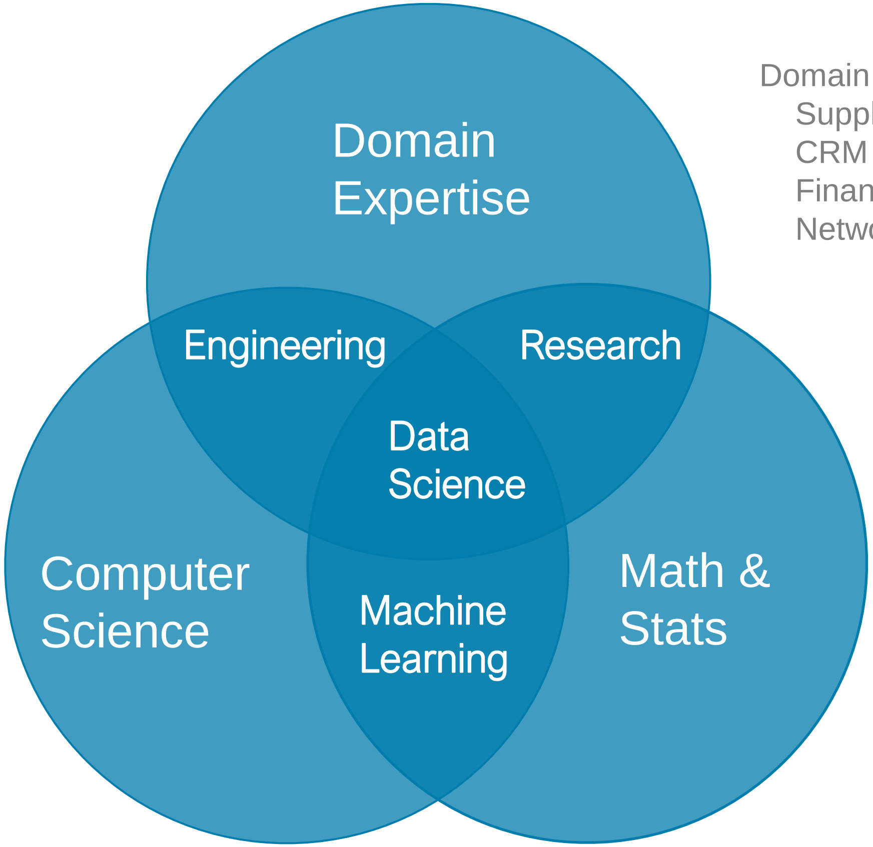
Outline

- **Data Science Overview**
- **Cloud Pak for Data Overview**
- **Lab Overview**

What is Data Science?

Domain Knowledge
- Supply Chain
- CRM
- Financials
- Networking

Scripting, SQL
Python, R Scala
Data Pipelines
Big Data/ Apache Spark



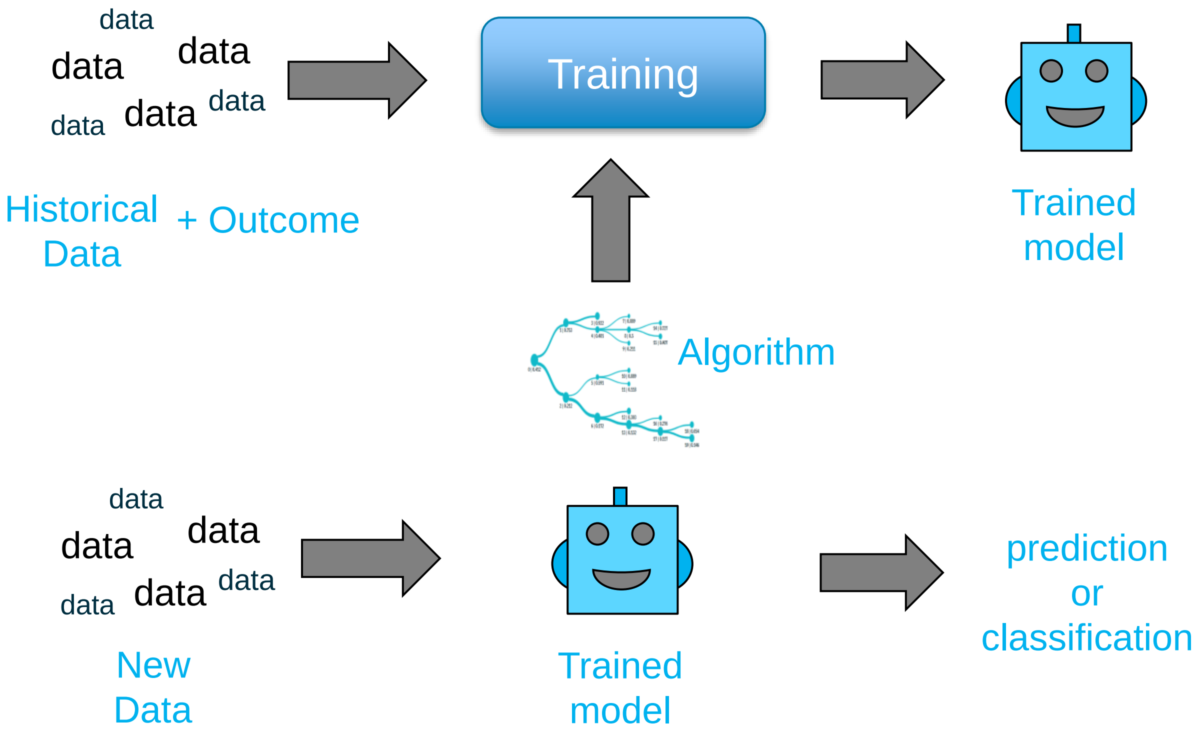
Mathematics
Computational

Data Science Projects Require Multiple Skills

Modified from Drew Conway's Venn Diagram

What is Machine Learning?

"Computers that learn without being explicitly programmed"



data data data data data data

Historical Data + Outcome

Training

Algorithm

Trained model

data data data data data data

New Data

Trained model

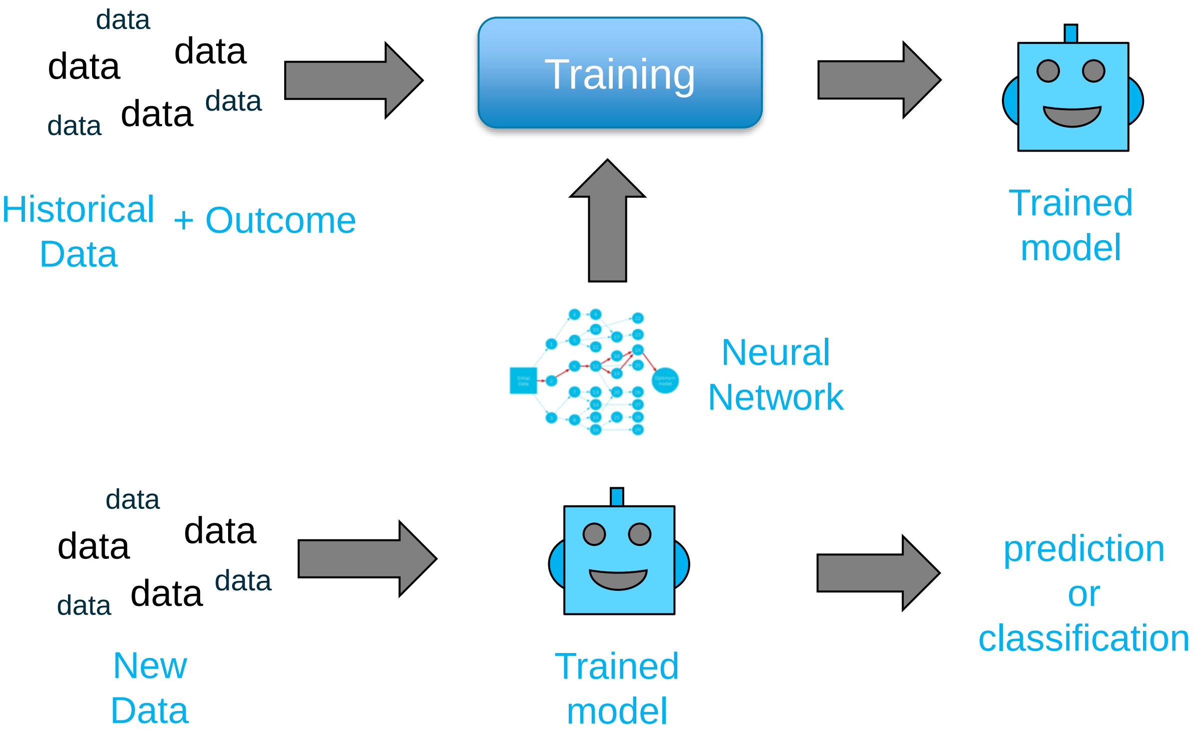
prediction or classification

1 A machine learning model is trained to recognize patterns in historical data

2 The model is then shown new data and asked to predict or classify it. If the patterns In the new data match the training data then the model makes accurate predictions.

What is Deep Learning?

"Computers that learn without being explicitly programmed"



data

data data

data data data

Historical Data + Outcome

Training

Trained model

Neural Network

data

data data

data data data

New Data

Trained model

prediction or classification

1 A machine learning model is trained to recognize patterns in historical data

2 The model is then shown new data and asked to predict or classify it. If the patterns In the new data match the training data then the model makes accurate predictions.

IBM takes an Enterprise Approach to Data Science

- Integrated Multi-modal platform
  - Use tool of choice and collaborate via project entities
  - Code/Click Options
  - All Analytics – Dashboard, Predictive, Prescriptive
  - All Data
  - Seamless user experience
- Hybrid Cloud
  - Cloud native architecture
  - Cloud agnostic – any vendor cloud or data center
  - Scalable data and analytic services
  - Flexibility to move data science to the data.
- Operationalize Machine Learning
  - Ease and flexibility of deployment at enterprise scale
  - Advanced model management capabilities.
  - Monitoring model performance
- Governance
  - Omnipresent, yet invisible – infused throughout
  - Data automatically integrated with governance capability for auto-discovery, catalog, and search subject to policies and rules
- Automate, Automate, Automate

Outline

- Data Science Overview
- Cloud Pak for Data Overview
- Lab Overview

Cloud Pak for Data Platform

An integrated platform of tools, services, data, and metadata that help companies or agencies accelerate their shift to be data-driven organizations.

Drive governance policy effectiveness while tracking how data is used and its value to the company

Data Steward

Make the insights immediately actionable and add intelligence to apps in straightforward manner

App Developer

Data Engineer

Easily build data pipelines that power dashboards and data platforms while ensuring high quality

Data Scientist

Access powerful tools to tease out the insights using advanced analytics

Business Analyst

Build dashboards and reports to visualize insights to the business

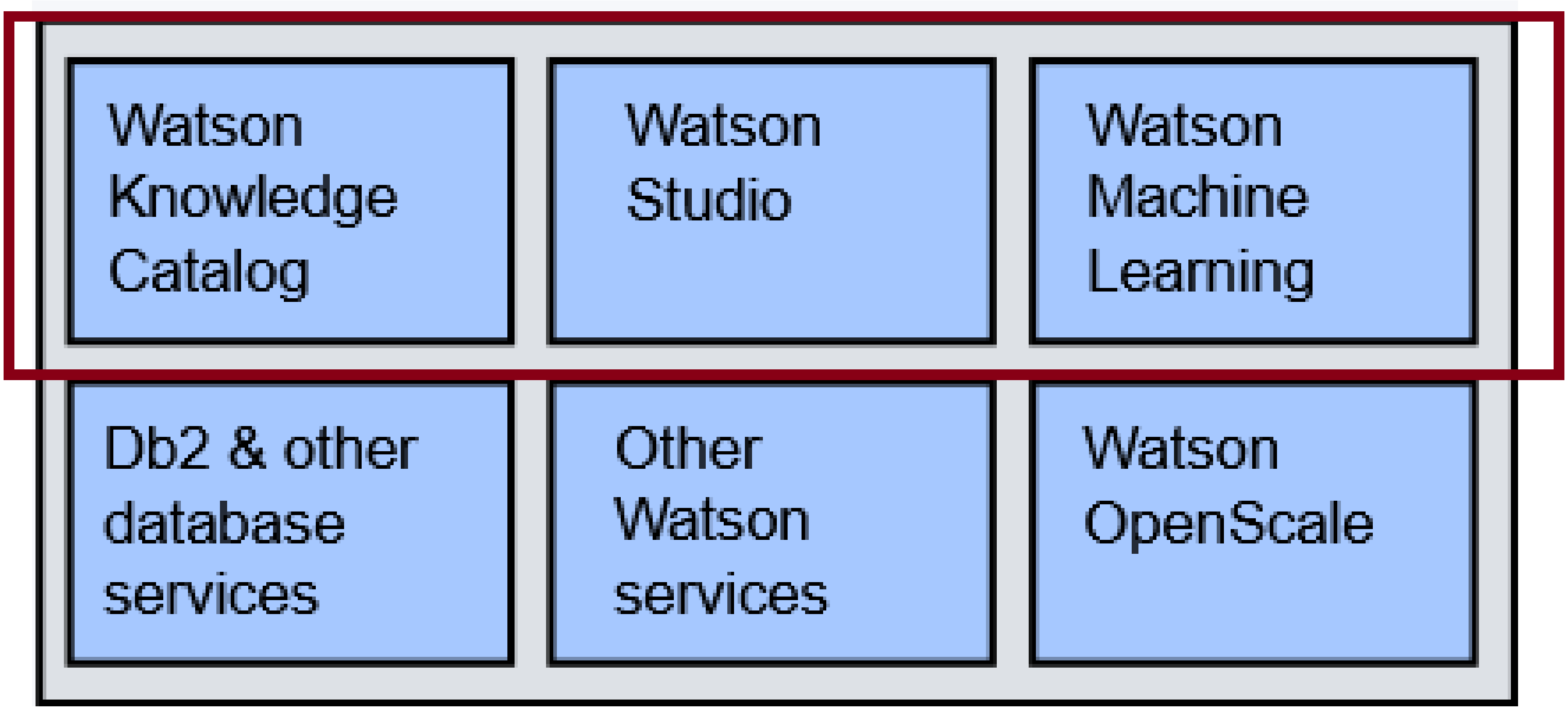
Cloud Pak for Data Deployment Options

- Cloud Pak for Data as a Service
  - Managed offering provided by IBM
  - Used for today’s labs

- Cloud Pak for Data
  - Available anywhere Red Hat OpenShift is supported
  - Public Clouds – IBM, Amazon Web Service, Microsoft Azure, Google Cloud
  - On-premise

- Cloud Pak for Data System
  - Pre-configured hardware
  - Same capabilities as Cloud Pak for Data
  - On-premise

Cloud Pak for Data as a Service



Watson Knowledge Catalog

Watson Studio

Watson Machine Learning

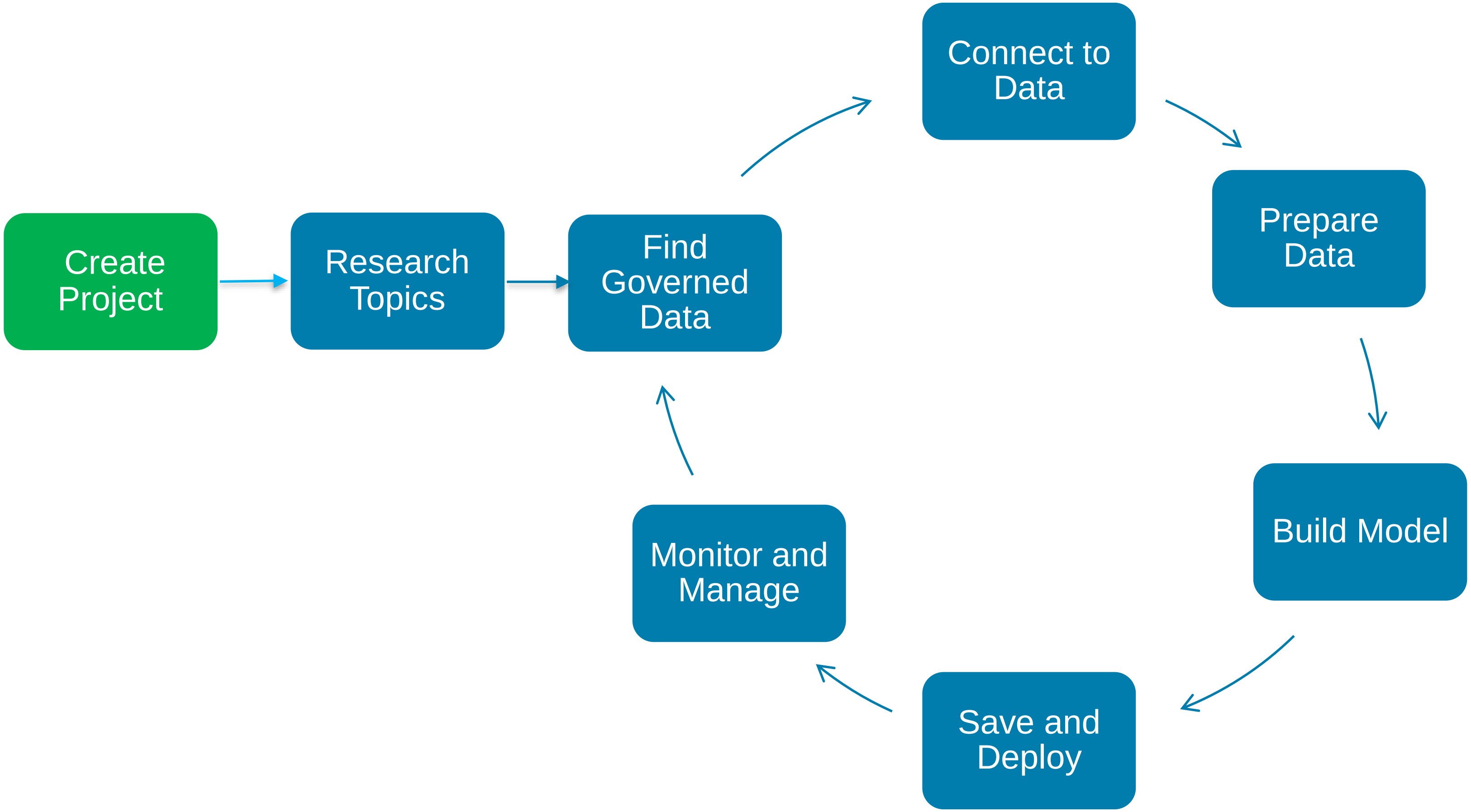
Db2 & other database services

Other Watson services

Watson OpenScale

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Create Project

Research Topics

Find Governed Data

Connect to Data

Prepare Data

Build Model

Save and Deploy

Monitor and Manage

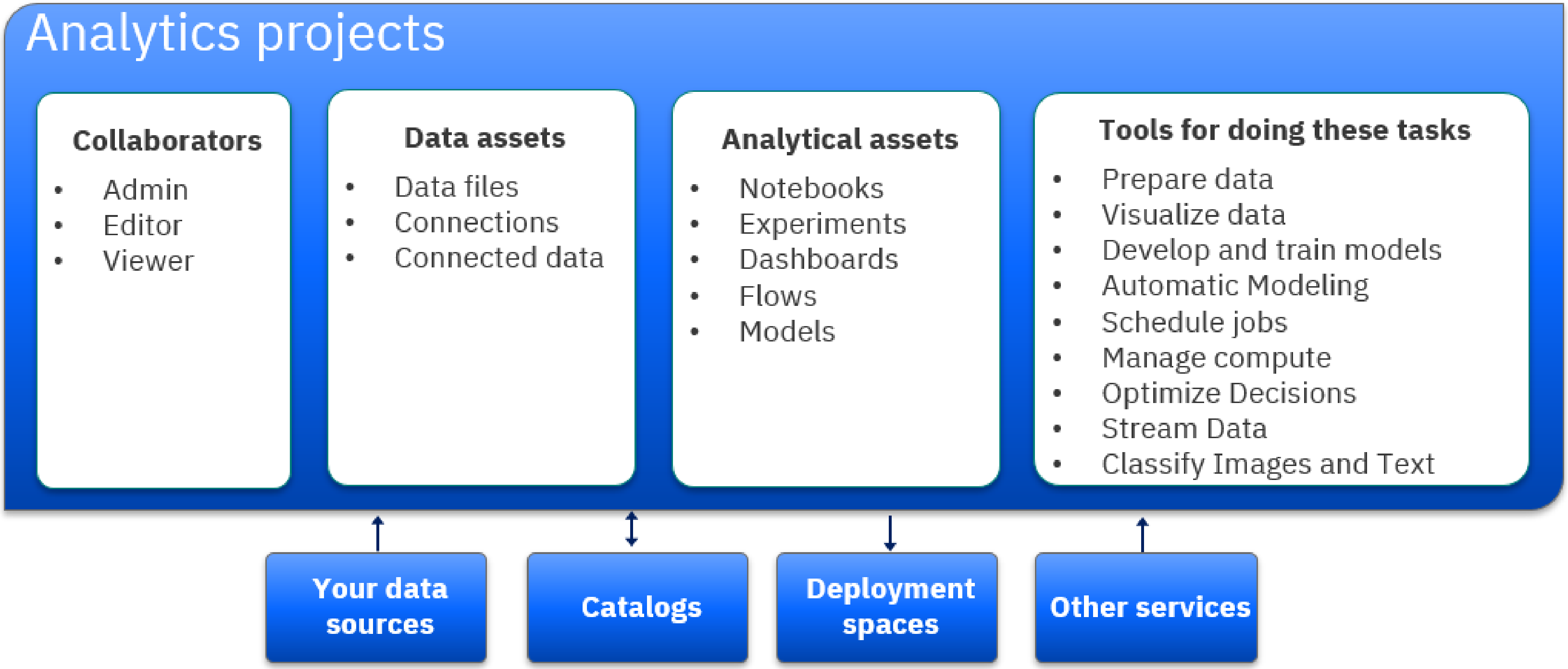
Watson Studio Project

Making Data Science a Team Sport

Create Project

Watson Studio provides the environment and tools to collaborate on business problems.

Watson Studio is centered around the Project. Data scientists and business analysts use projects to organize resources and analyze data with various tools.



Analytics projects

Collaborators

- Admin
- Editor
- Viewer

Data assets

- Data files
- Connections
- Connected data

Analytical assets

- Notebooks
- Experiments
- Dashboards
- Flows
- Models

Tools for doing these tasks

- Prepare data
- Visualize data
- Develop and train models
- Automatic Modeling
- Schedule jobs
- Manage compute
- Optimize Decisions
- Stream Data
- Classify Images and Text

Your data sources

Catalogs

Deployment spaces

Other services

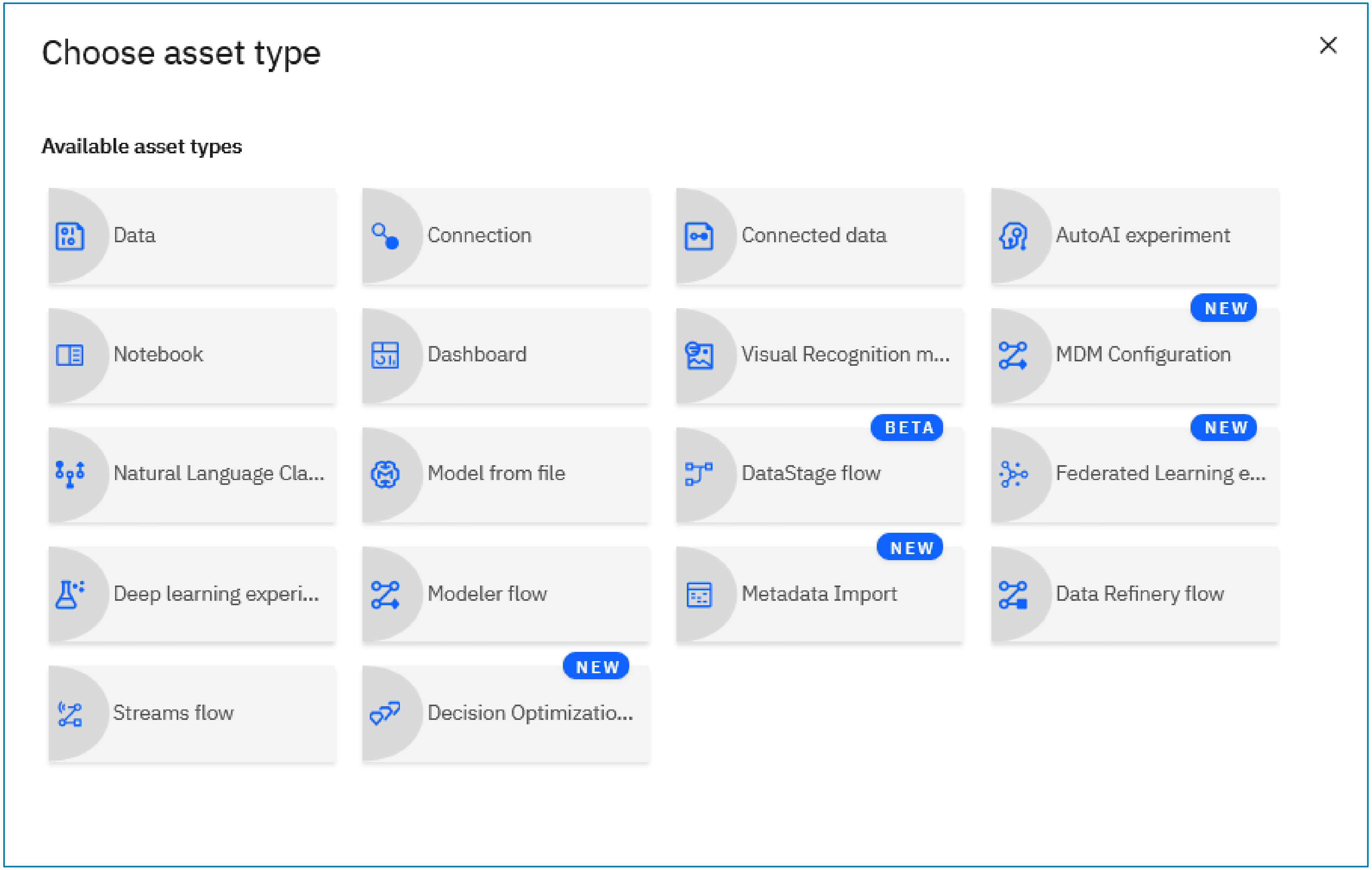
Watson Studio Project Features

Making Data Science a Team Sport

Create Project

- Organizes resources to achieve a particular data analysis goal
- Support role-based collaboration (Admin, Editor, Viewer)
- Assets from all IDEs can be included in one Watson Studio project: notebooks, data sources, flows, models, etc.
- Export/Import Projects

Add to Project



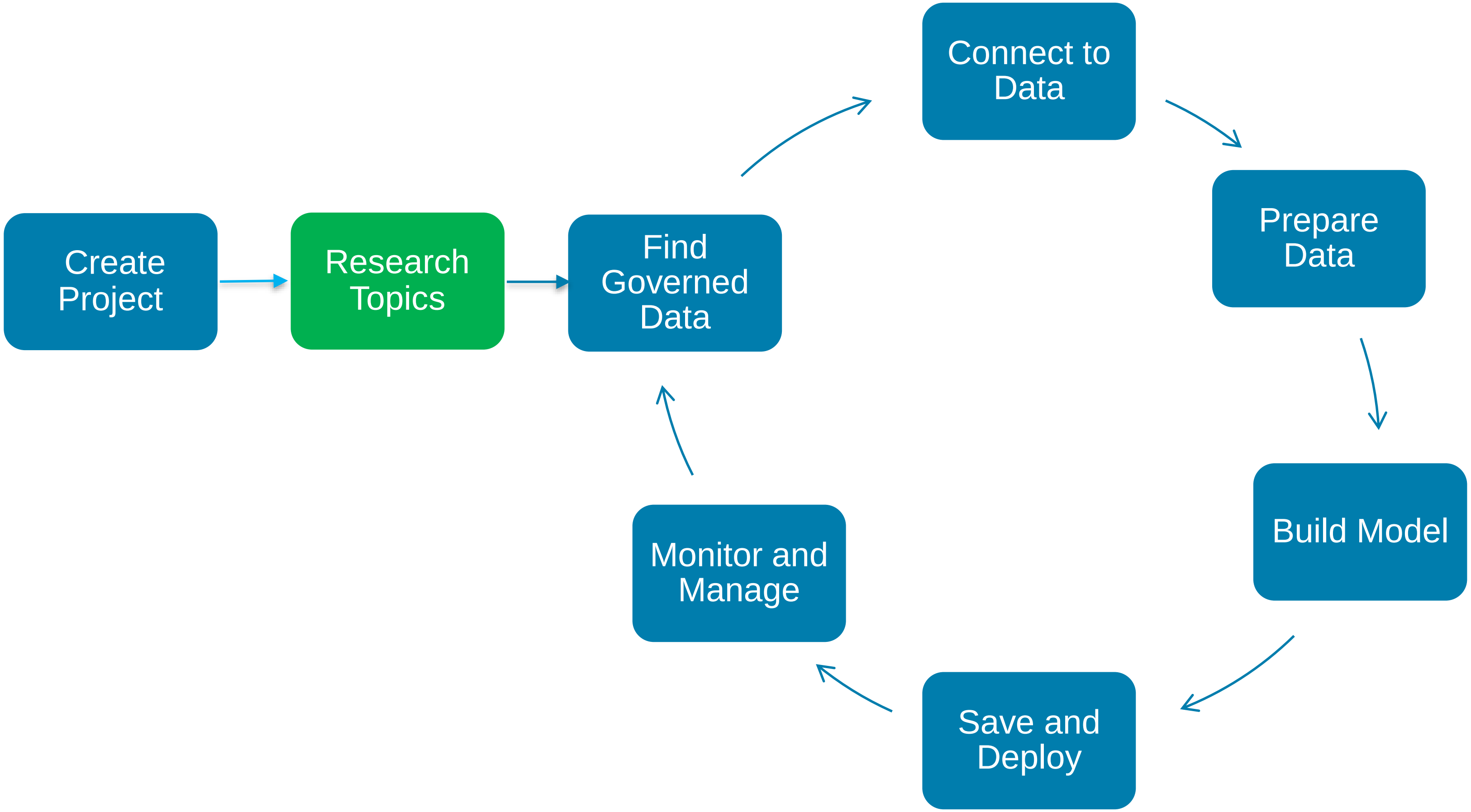
Choose asset type

Available asset types

- Data
- Connection
- Connected data
- AutoAI experiment
- Notebook
- Dashboard
- Visual Recognition m...
- NEW MDM Configuration
- Natural Language Cla...
- Model from file
- BETA DataStage flow
- NEW Federated Learning e...
- Deep learning experi...
- Modeler flow
- NEW Metadata Import
- Data Refinery flow
- Streams flow
- NEW Decision Optimizatio...

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Watson Studio Gallery

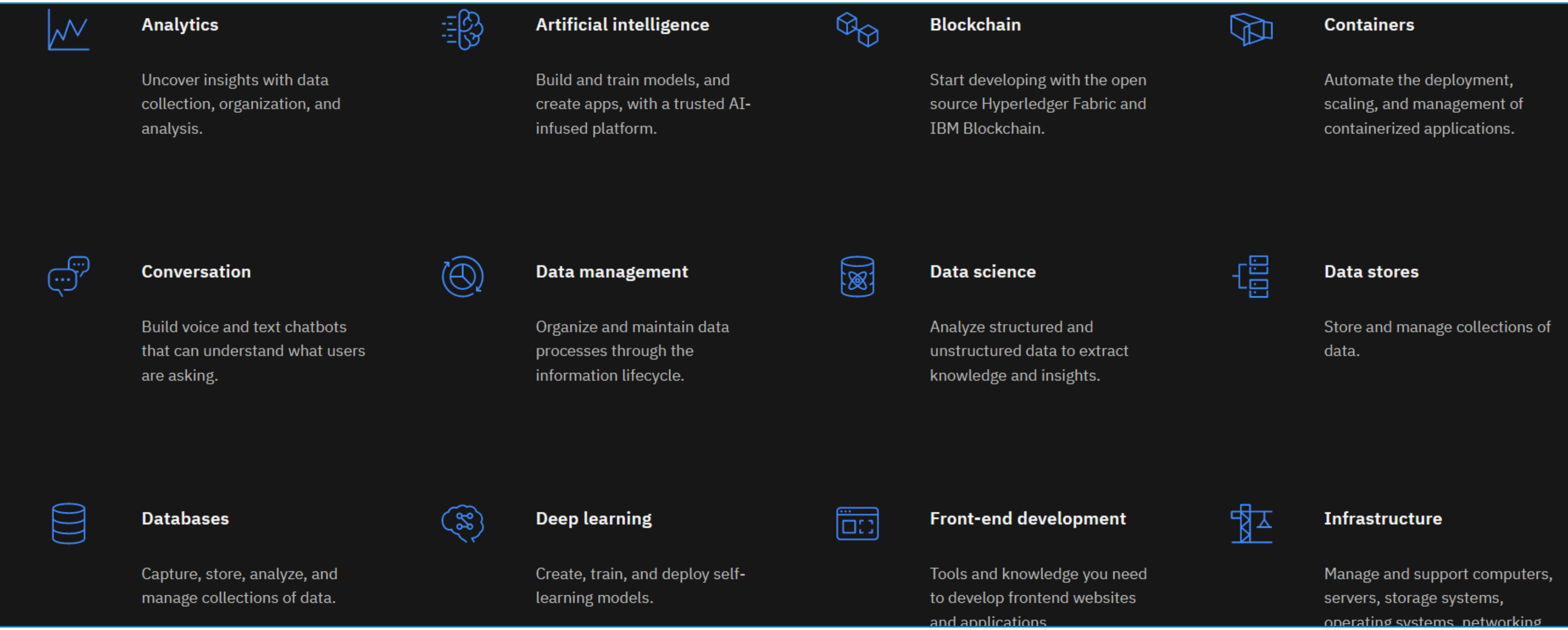
Built-in learning to get started

Research Topics

- The Gallery includes sample projects, notebooks, and data sets
- Copy notebooks or Data Sets into projects
- Instantiate sample projects
- Continuously updated in IBM’s Cloud Pak for Data as a Service

developer.ibm.com/technologies

Research Topics



Analytics

Uncover insights with data collection, organization, and analysis.

Artificial intelligence

Build and train models, and create apps, with a trusted AI-infused platform.

Blockchain

Start developing with the open source Hyperledger Fabric and IBM Blockchain.

Containers

Automate the deployment, scaling, and management of containerized applications.

Conversation

Build voice and text chatbots that can understand what users are asking.

Data management

Organize and maintain data processes through the information lifecycle.

Data science

Analyze structured and unstructured data to extract knowledge and insights.

Data stores

Store and manage collections of data.

Databases

Capture, store, analyze, and manage collections of data.

Deep learning

Create, train, and deploy self-learning models.

Front-end development

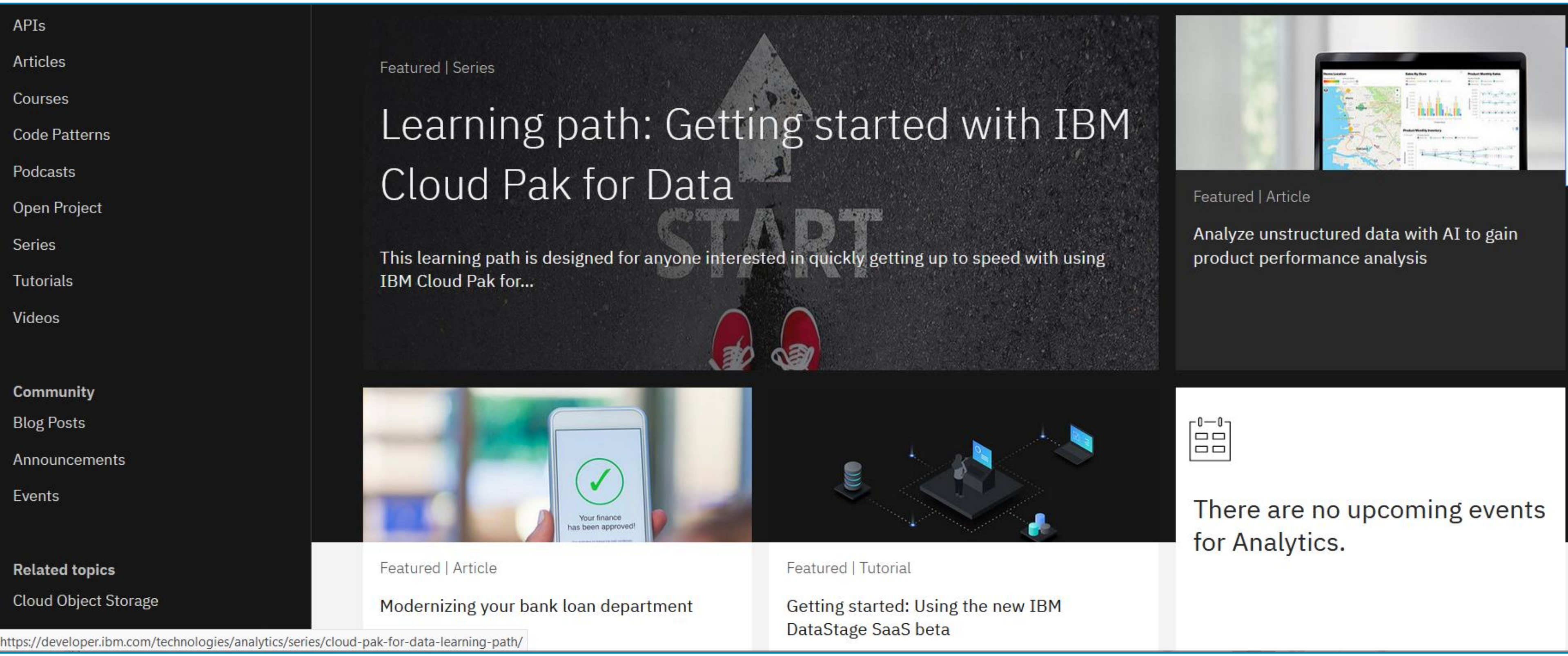
Tools and knowledge you need to develop frontend websites and applications

Infrastructure

Manage and support computers, servers, storage systems, operating systems, networking

developer.ibm.com

Research Topics



APIs

Articles

Courses

Code Patterns

Podcasts

Open Project

Series

Tutorials

Videos

Community

Blog Posts

Announcements

Events

Related topics

Cloud Object Storage

Featured | Series

Learning path: Getting started with IBM Cloud Pak for Data

This learning path is designed for anyone interested in quickly getting up to speed with using IBM Cloud Pak for...

Featured | Article

Analyze unstructured data with AI to gain product performance analysis

Featured | Article

Modernizing your bank loan department

Featured | Tutorial

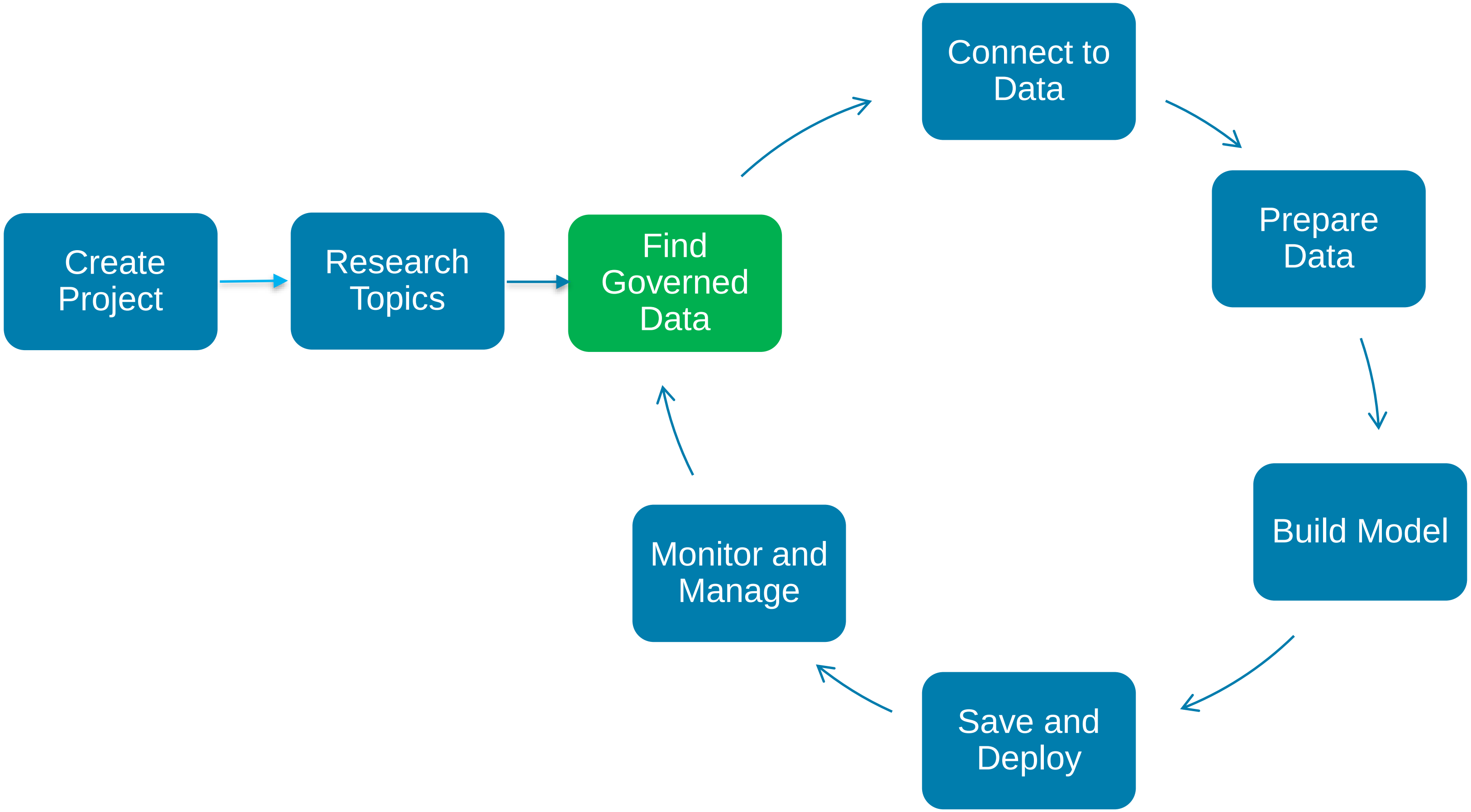
Getting started: Using the new IBM DataStage SaaS beta

There are no upcoming events for Analytics.

https://developer.ibm.com/technologies/analytics/series/cloud-pak-for-data-learning-path/

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Watson Knowledge Catalog Features

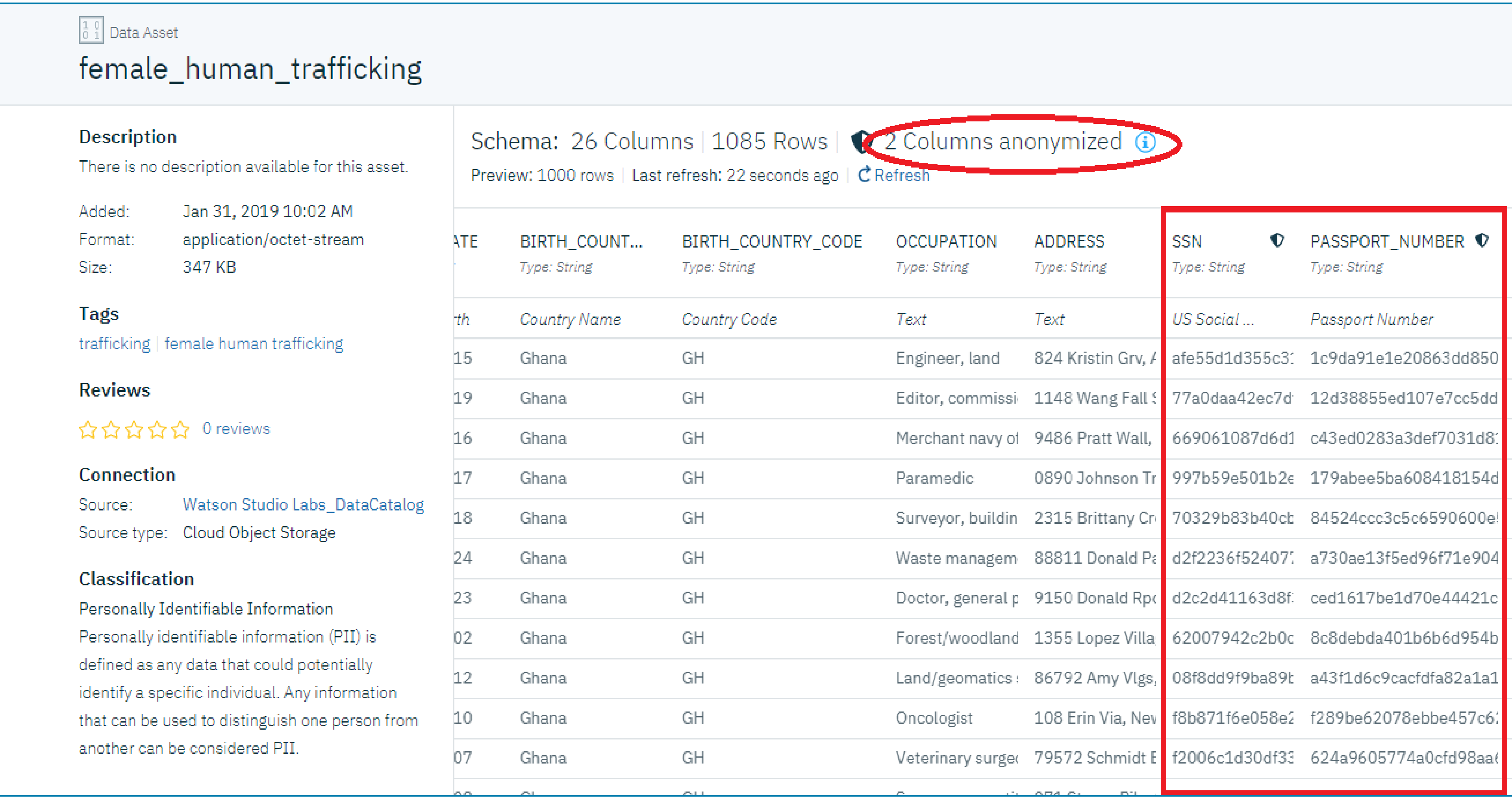
Unlock tribal knowledge and unleash knowledge workers

Find Governed Data

- **Find** data (structured, unstructured) and AI assets (e.g., ML/DL models, notebooks, Watson Data Kits) in the **Knowledge Catalog** with intelligent search and giving the right access to the right users.
- Discover assets, profiling, classification
- Policy, rule authoring
- Policy, rule enforcement
- Asset Usage Statistics

Watson Knowledge Catalog Features

Find Governed Data



Data Asset

female_human_trafficking

Description

There is no description available for this asset.

Added: Jan 31, 2019 10:02 AM
Format: application/octet-stream
Size: 347 KB

Tags

trafficking | female human trafficking

Reviews

☆☆☆☆☆ 0 reviews

Connection

Source: Watson Studio Labs_DataCatalog
Source type: Cloud Object Storage

Classification

Personally Identifiable Information

Personally identifiable information (PII) is defined as any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another can be considered PII.

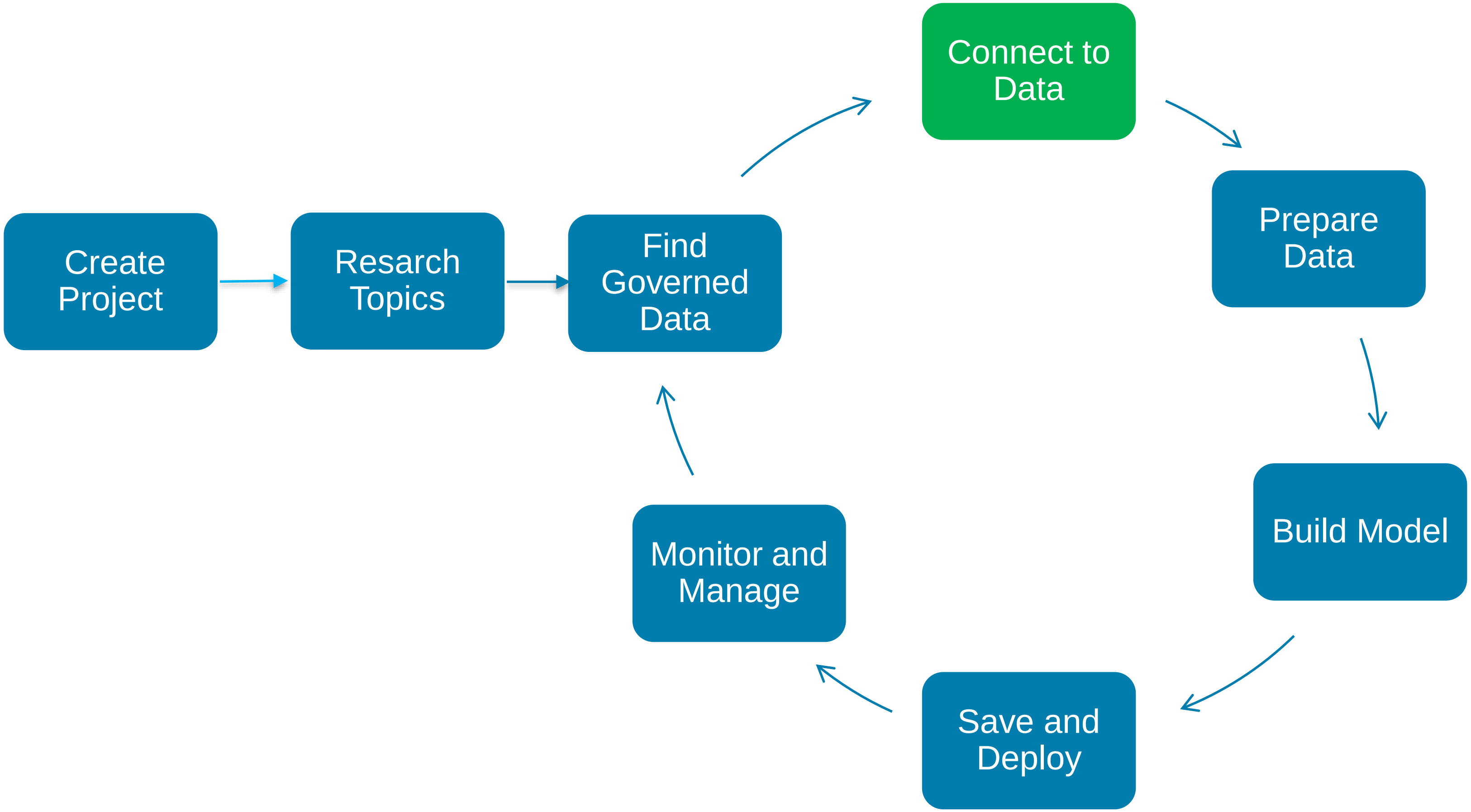
Schema: 26 Columns | 1085 Rows | 2 Columns anonymized

Preview: 1000 rows | Last refresh: 22 seconds ago | Refresh

ATE	BIRTH_COUNT... Type: String	BIRTH_COUNTRY_CODE Type: String	OCCUPATION Type: String	ADDRESS Type: String	SSN Type: String	PASSPORT_NUMBER Type: String
th	*Country Name*	*Country Code*	*Text*	*Text*	*US Social ...*	*Passport Number*
15	Ghana	GH	Engineer, land	824 Kristin Grv, A	afe55d1d355c3:	1c9da91e1e20863dd850
19	Ghana	GH	Editor, commissi	1148 Wang Fall S	77a0daa42ec7d	12d38855ed107e7cc5dd
16	Ghana	GH	Merchant navy of	9486 Pratt Wall,	669061087d6d1	c43ed0283a3def7031d8:
17	Ghana	GH	Paramedic	0890 Johnson Tr	997b59e501b2e	179abee5ba608418154d
18	Ghana	GH	Surveyor, buildin	2315 Brittany Cr	70329b83b40cb	84524ccc3c5c6590600e!
24	Ghana	GH	Waste managem	88811 Donald Pa	d2f2236f52407	a730ae13f5ed96f71e904
23	Ghana	GH	Doctor, general p	9150 Donald Rpc	d2c2d41163d8f:	ced1617be1d70e44421c
02	Ghana	GH	Forest/woodland	1355 Lopez Villa	62007942c2b0c	8c8debda401b6b6d954b
12	Ghana	GH	Land/geomatics :	86792 Amy Vlgs,	08f8dd9f9ba89b	a43f1d6c9cacfdfa82a1a1
10	Ghana	GH	Oncologist	108 Erin Via, New	f8b871f6e058e2	f289be62078ebbe457c6:
07	Ghana	GH	Veterinary surge	79572 Schmidt E	f2006c1d30df33	624a9605774a0cfd98aa

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



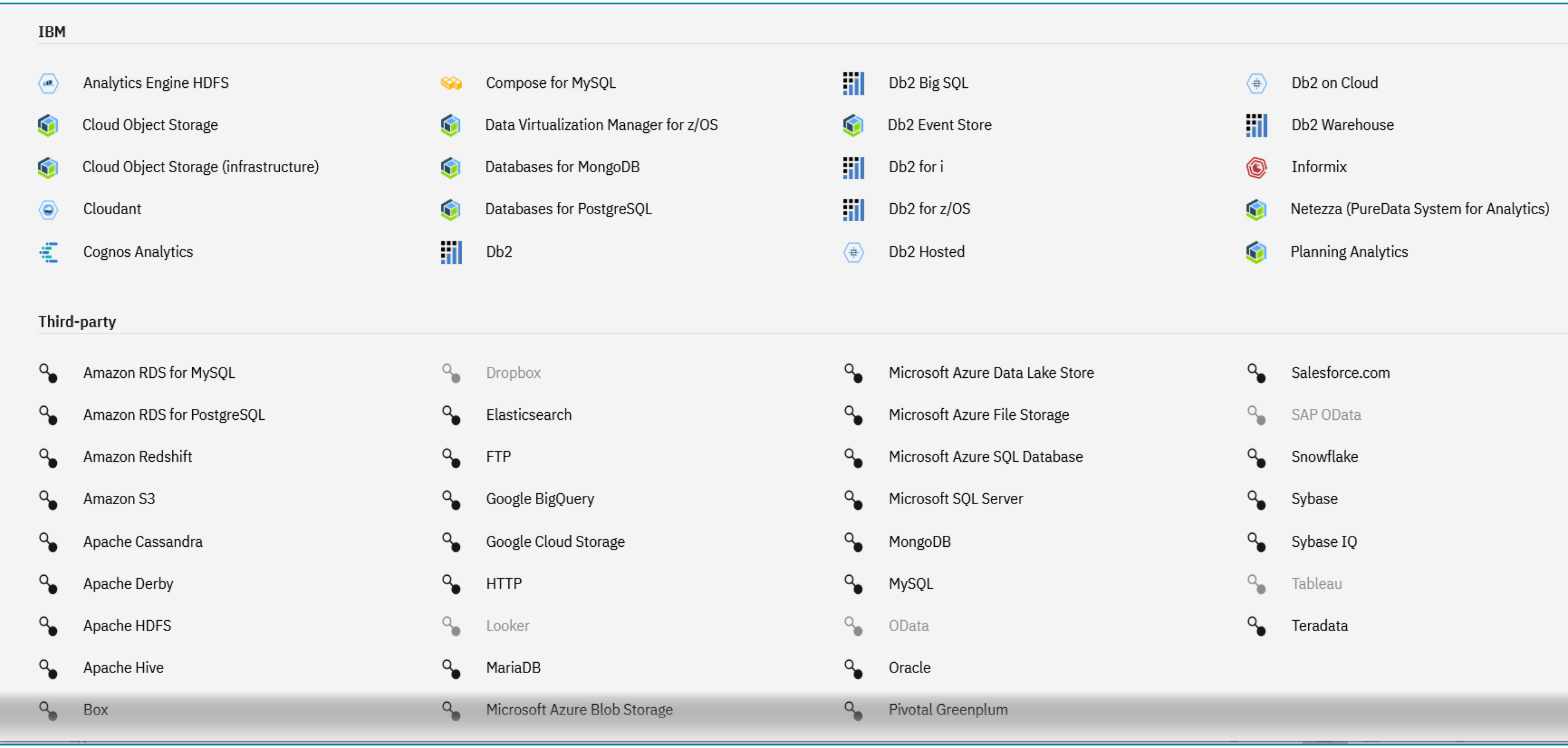
Watson Studio Connection Features

Connect to Data

- Upload files
- Connectors to Structured and Unstructured, On-prem and Cloud data sources.
- Wizard based connection definition and code generation

Connection Options

Connect to Data



IBM

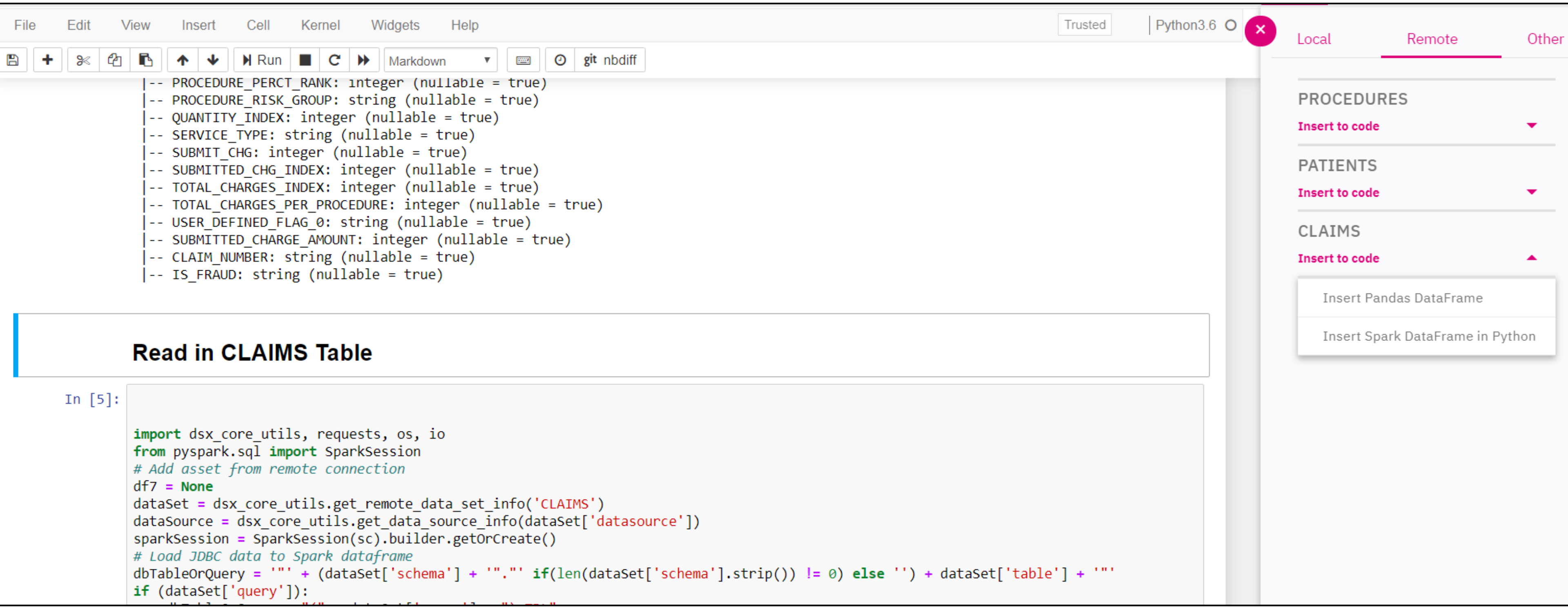
- Analytics Engine HDFS
- Cloud Object Storage
- Cloud Object Storage (infrastructure)
- Cloudant
- Cognos Analytics
- Compose for MySQL
- Data Virtualization Manager for z/OS
- Databases for MongoDB
- Databases for PostgreSQL
- Db2
- Db2 Big SQL
- Db2 Event Store
- Db2 for i
- Db2 for z/OS
- Db2 Hosted
- Db2 on Cloud
- Db2 Warehouse
- Informix
- Netezza (PureData System for Analytics)
- Planning Analytics

Third-party

- Amazon RDS for MySQL
- Amazon RDS for PostgreSQL
- Amazon Redshift
- Amazon S3
- Apache Cassandra
- Apache Derby
- Apache HDFS
- Apache Hive
- Box
- Dropbox
- Elasticsearch
- FTP
- Google BigQuery
- Google Cloud Storage
- HTTP
- Looker
- MariaDB
- Microsoft Azure Blob Storage
- Microsoft Azure Data Lake Store
- Microsoft Azure File Storage
- Microsoft Azure SQL Database
- Microsoft SQL Server
- MongoDB
- MySQL
- OData
- Oracle
- Pivotal Greenplum
- Salesforce.com
- SAP OData
- Snowflake
- Sybase
- Sybase IQ
- Tableau
- Teradata

Notebook Screenshot

Connect to Data



```
 |-- PROCEDURE_PERCT_RANK: integer (nullable = true)
 |-- PROCEDURE_RISK_GROUP: string (nullable = true)
 |-- QUANTITY_INDEX: integer (nullable = true)
 |-- SERVICE_TYPE: string (nullable = true)
 |-- SUBMIT_CHG: integer (nullable = true)
 |-- SUBMITTED_CHG_INDEX: integer (nullable = true)
 |-- TOTAL_CHARGES_INDEX: integer (nullable = true)
 |-- TOTAL_CHARGES_PER_PROCEDURE: integer (nullable = true)
 |-- USER_DEFINED_FLAG_0: string (nullable = true)
 |-- SUBMITTED_CHARGE_AMOUNT: integer (nullable = true)
 |-- CLAIM_NUMBER: string (nullable = true)
 |-- IS_FRAUD: string (nullable = true)
```

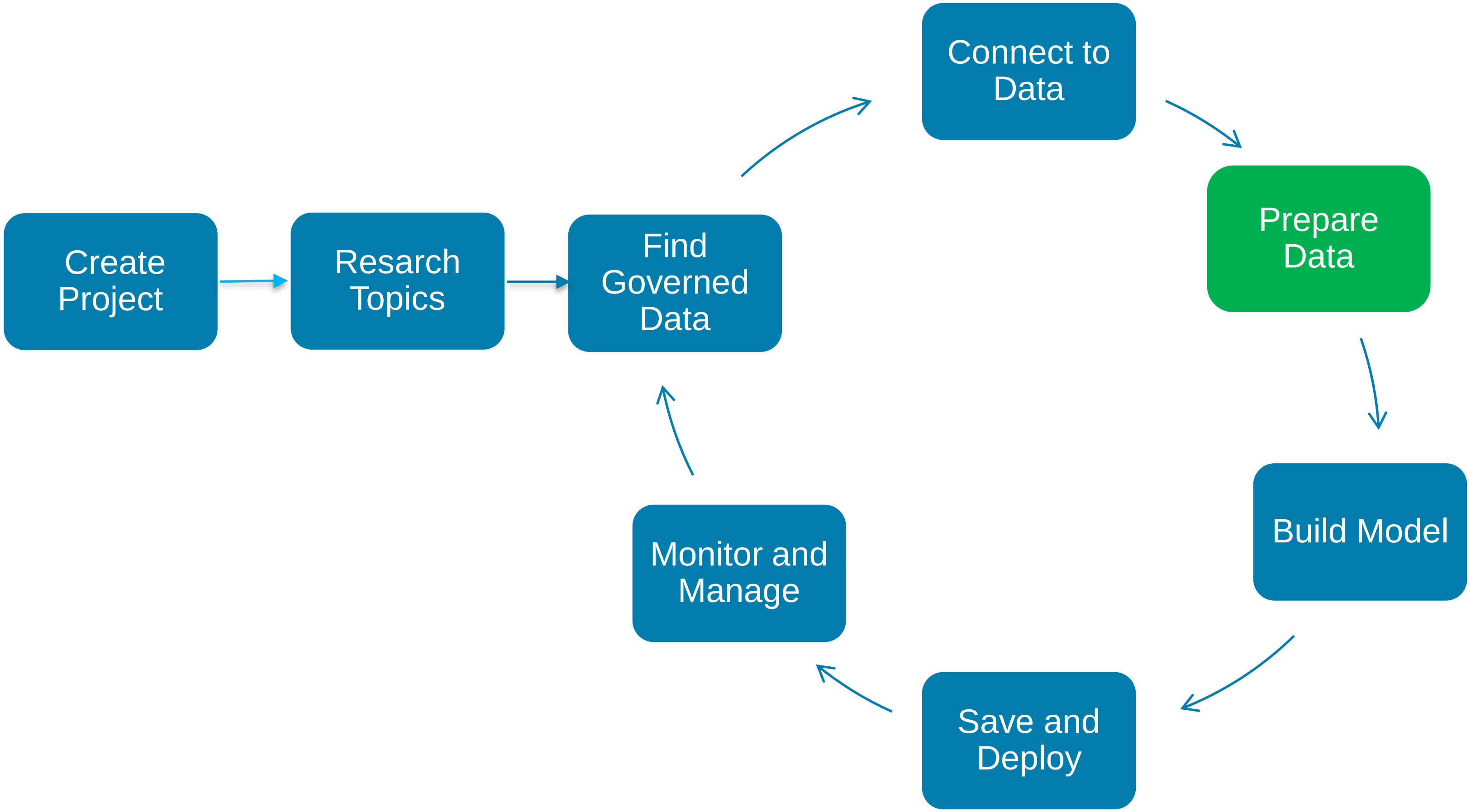
Read in CLAIMS Table

In [5]:

```
import dsx_core_utils, requests, os, io
from pyspark.sql import SparkSession
# Add asset from remote connection
df7 = None
dataSet = dsx_core_utils.get_remote_data_set_info('CLAIMS')
dataSource = dsx_core_utils.get_data_source_info(dataSet['datasource'])
sparkSession = SparkSession(sc).builder.getOrCreate()
# Load JDBC data to Spark dataframe
dbTableOrQuery = '"' + (dataSet['schema'] + '"."' if(len(dataSet['schema'].strip()) != 0) else '') + dataSet['table'] + '"'
if (dataSet['query']):
```

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Watson Studio Data Refinery Features

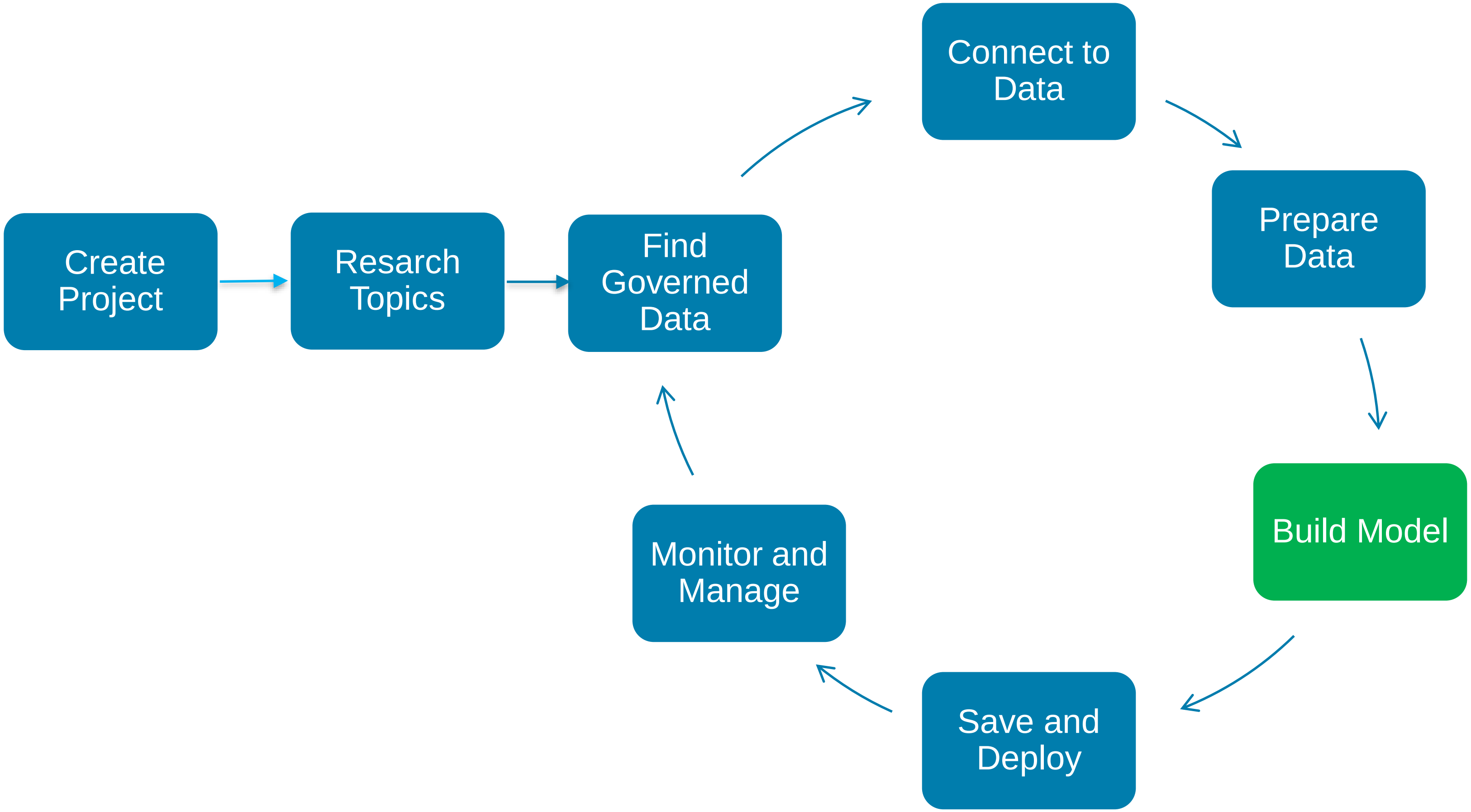
Making Data fit for use

Prepare Data

- Data Refinery tool to profile, visualize, and shape data.
- Create data preparation pipelines via point and click capability on subset of data
  - ✓ Cleanse the data: fixing or removing data that is incorrect, incomplete, improperly formatted, or duplicated
  - ✓ Shape the data: customize data by filtering, sorting, combining, or removing columns, and performing operations
- Run the pipeline on all the data
  - Manually (on demand)
  - Automated (scheduled)

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Watson Studio Model Building Features

The best of open source and IBM Watson tools to create start-of-the-art data products

Build Model

Open Source Tools

- Jupyter Notebooks**
- RStudio and Shiny
- Libraries- scikit-learn, XGBoost**, Spark**, TensorFlow, Keras, PyTorch

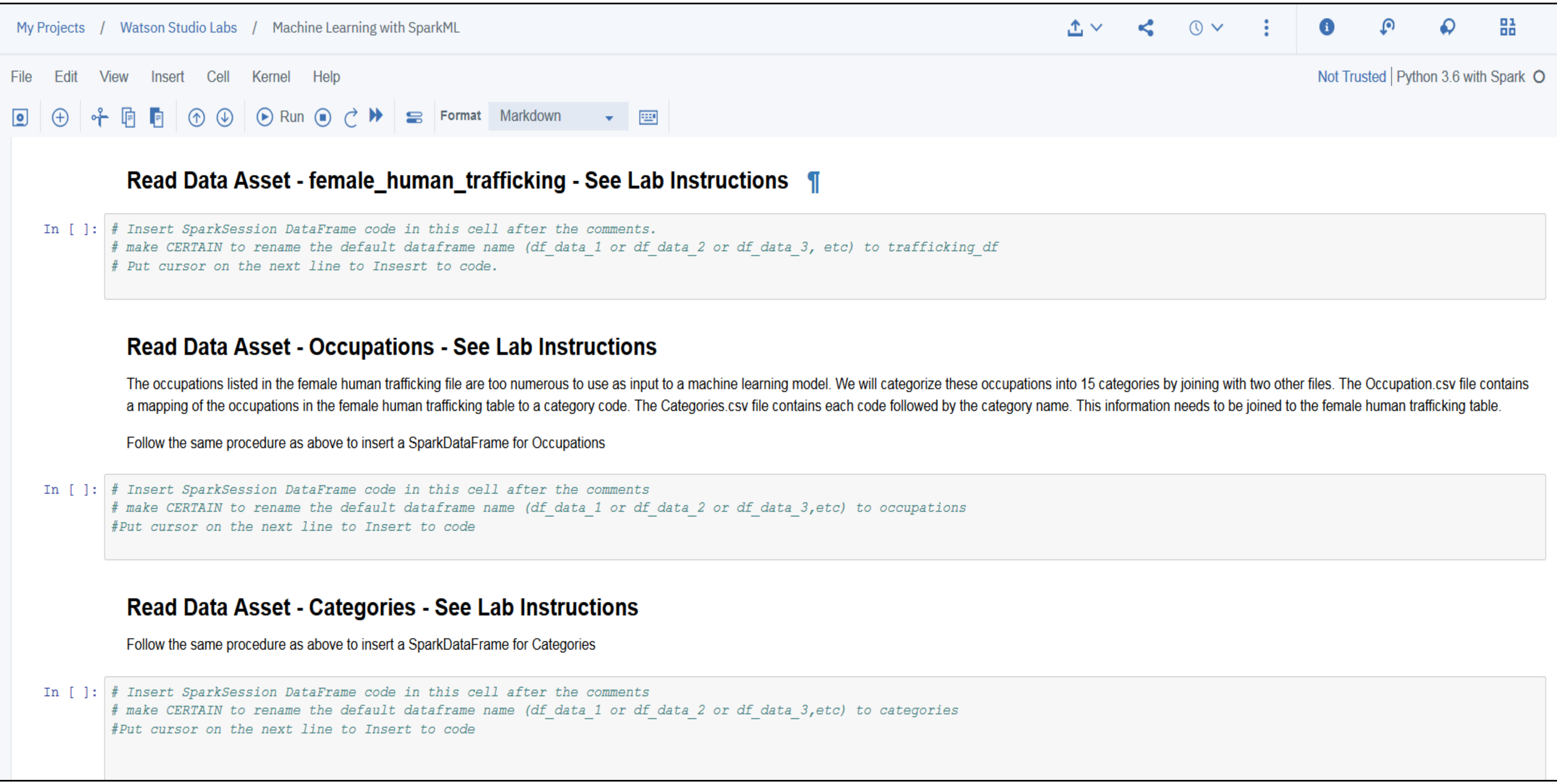
IBM Tools

- AutoAI **
- SPSS Modeler**
- Experiment Builder
- Natural Language Classifier Model
- Visual Recognition Model
- IBM Streams Designer
- IBM Decision Optimization **
- Federated Learning (Beta)

** in hands-on labs

Jupyter Notebook

Build Model



Read Data Asset - female_human_trafficking - See Lab Instructions ¶

```
In [ ]: # Insert SparkSession DataFrame code in this cell after the comments.
        # make CERTAIN to rename the default dataframe name (df_data_1 or df_data_2 or df_data_3, etc) to trafficking_df
        # Put cursor on the next line to Insesrt to code.
```

Read Data Asset - Occupations - See Lab Instructions

The occupations listed in the female human trafficking file are too numerous to use as input to a machine learning model. We will categorize these occupations into 15 categories by joining with two other files. The Occupation.csv file contains a mapping of the occupations in the female human trafficking table to a category code. The Categories.csv file contains each code followed by the category name. This information needs to be joined to the female human trafficking table.

Follow the same procedure as above to insert a SparkDataFrame for Occupations

```
In [ ]: # Insert SparkSession DataFrame code in this cell after the comments
        # make CERTAIN to rename the default dataframe name (df_data_1 or df_data_2 or df_data_3,etc) to occupations
        #Put cursor on the next line to Insert to code
```

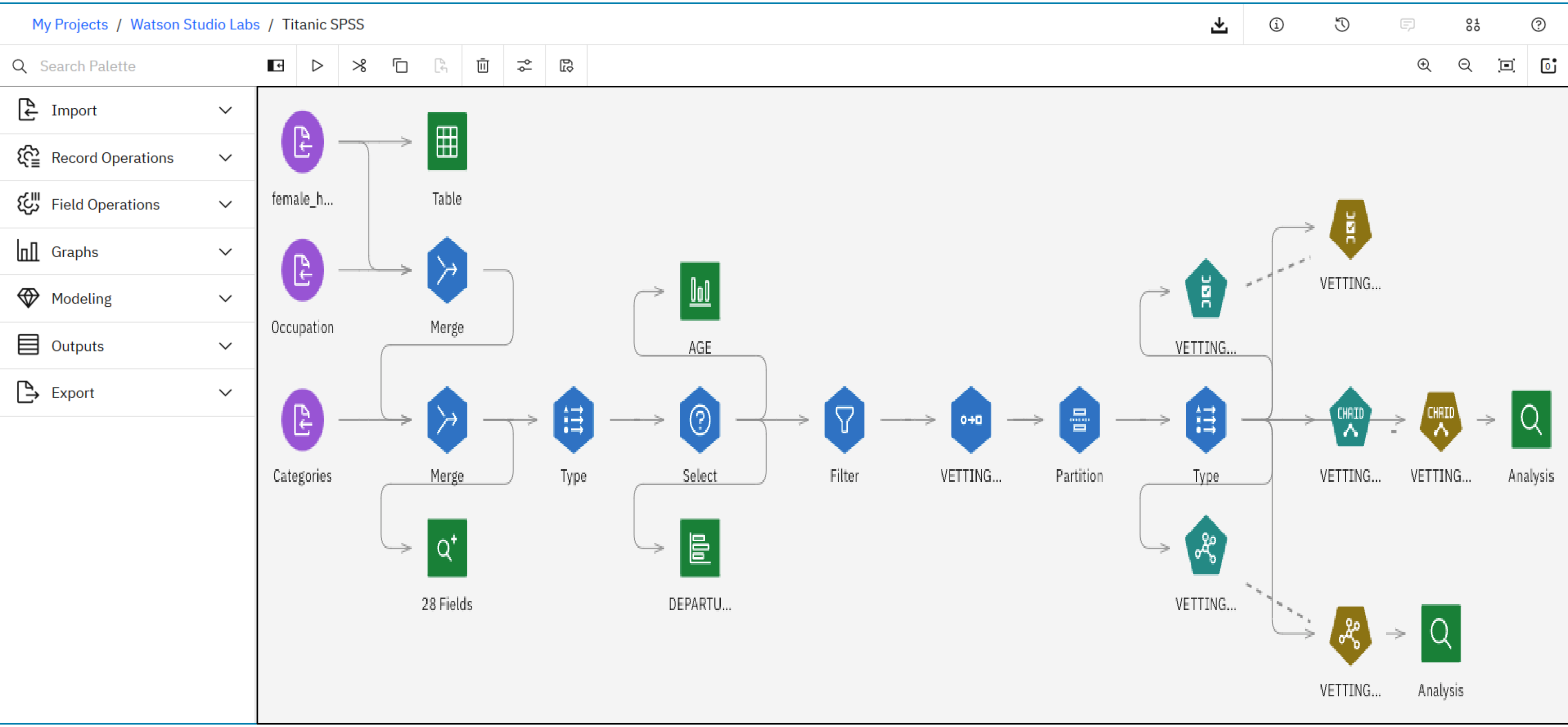
Read Data Asset - Categories - See Lab Instructions

Follow the same procedure as above to insert a SparkDataFrame for Categories

```
In [ ]: # Insert SparkSession DataFrame code in this cell after the comments
        # make CERTAIN to rename the default dataframe name (df_data_1 or df_data_2 or df_data_3,etc) to categories
        #Put cursor on the next line to Insert to code
```

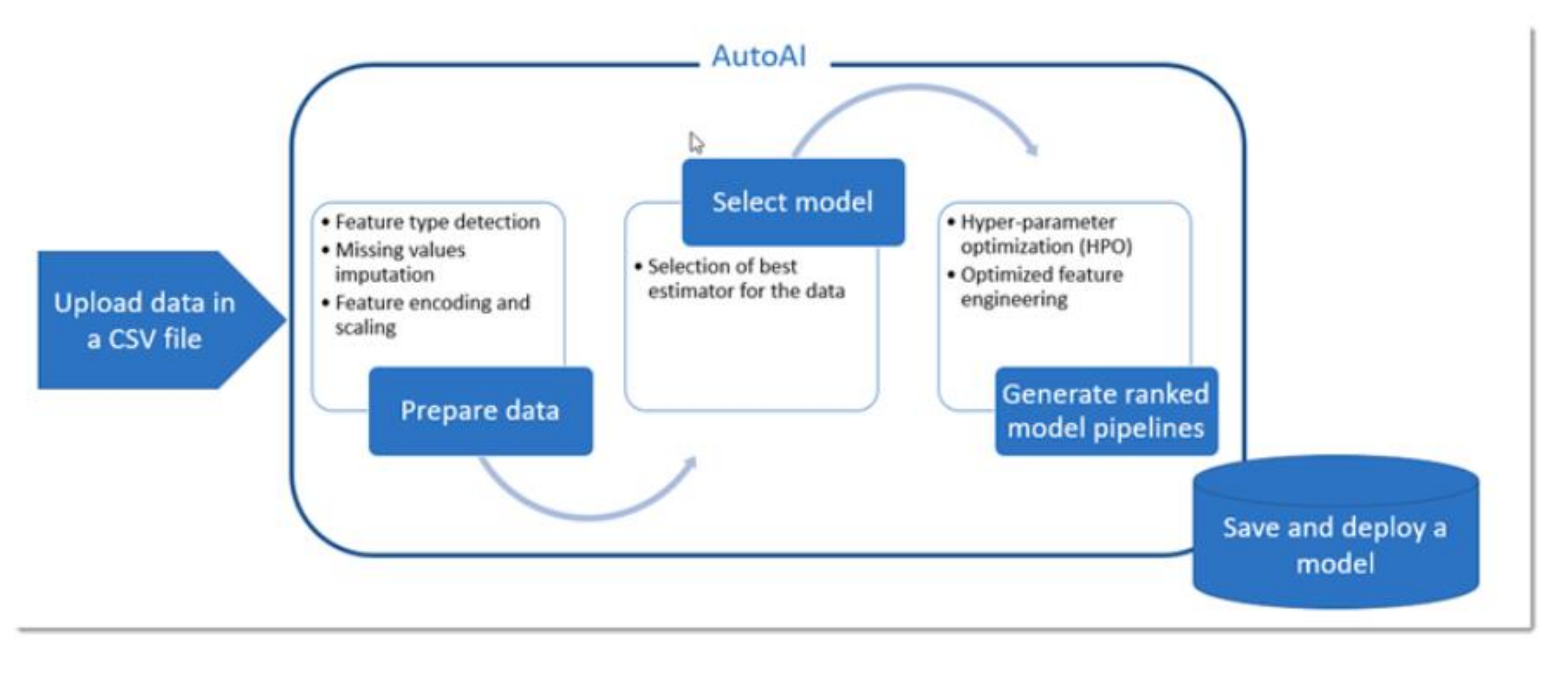
SPSS Modeler

Build Model



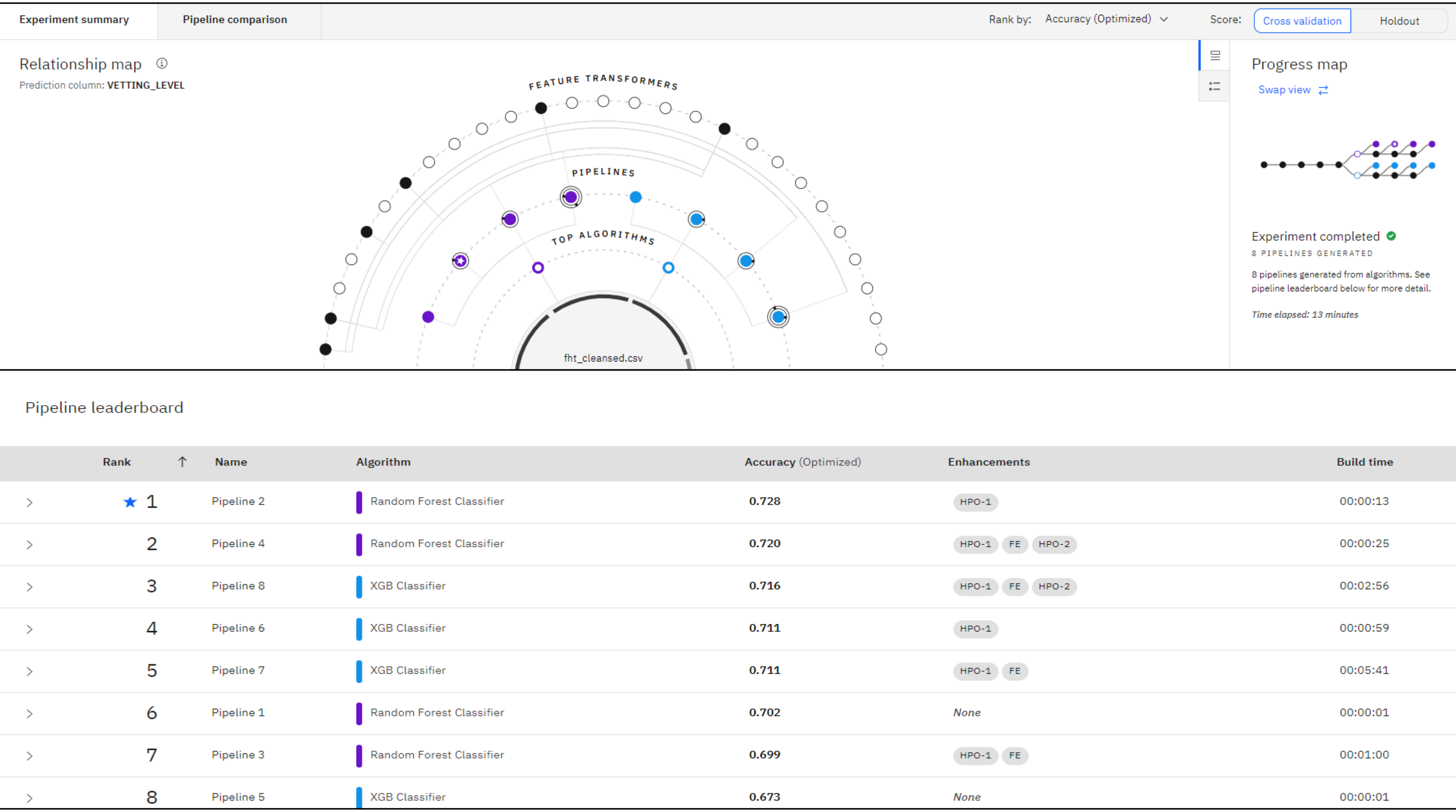
AutoAI

Build Model


AutoAI
Upload data in a CSV file
• Feature type detection
• Missing values imputation
• Feature encoding and scaling
Prepare data
Select model
• Selection of best estimator for the data
• Hyper-parameter optimization (HPO)
• Optimized feature engineering
Generate ranked model pipelines
Save and deploy a model

AutoAI

Build Model



Relationship map

Prediction column: VETTING_LEVEL

Progress map

Experiment completed

8 PIPELINES GENERATED

8 pipelines generated from algorithms. See pipeline leaderboard below for more detail.

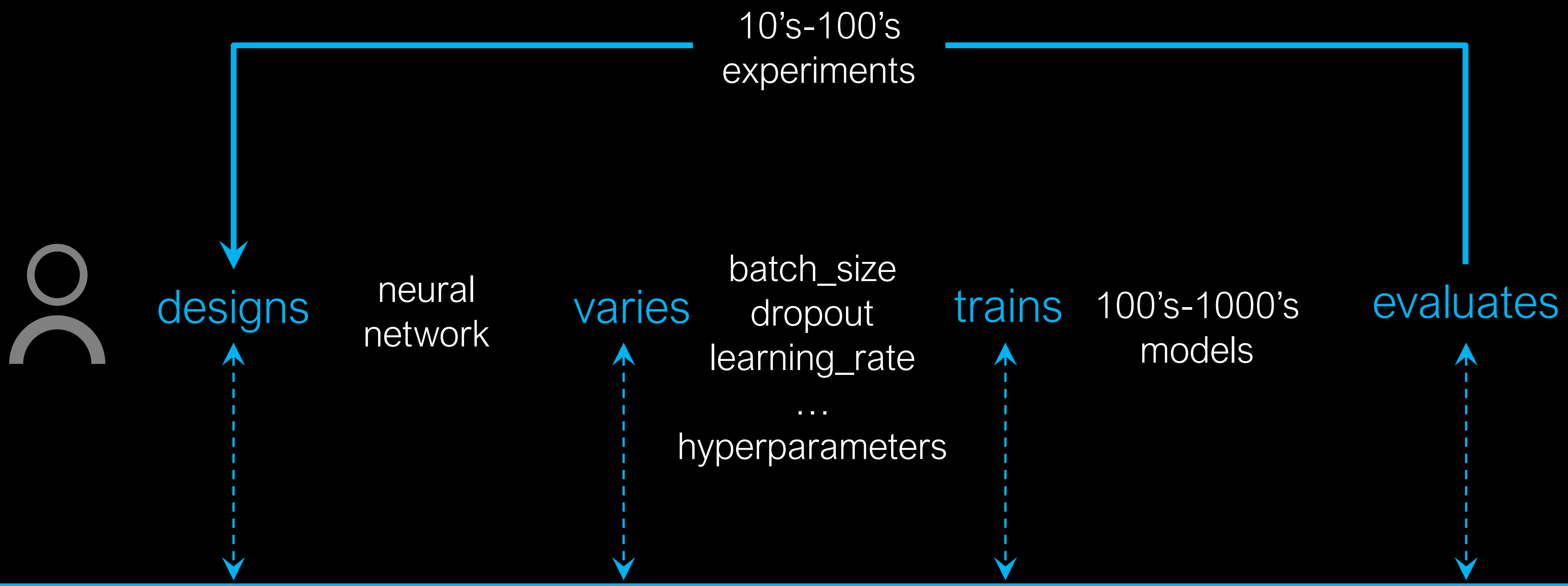
Time elapsed: 13 minutes

Pipeline leaderboard

Rank	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
★ 1	Pipeline 2	Random Forest Classifier	**0.728**	HPO-1	00:00:13
2	Pipeline 4	Random Forest Classifier	**0.720**	HPO-1 FE HPO-2	00:00:25
3	Pipeline 8	XGB Classifier	**0.716**	HPO-1 FE HPO-2	00:02:56
4	Pipeline 6	XGB Classifier	**0.711**	HPO-1	00:00:59
5	Pipeline 7	XGB Classifier	**0.711**	HPO-1 FE	00:05:41
6	Pipeline 1	Random Forest Classifier	**0.702**	*None*	00:00:01
7	Pipeline 3	Random Forest Classifier	**0.699**	HPO-1 FE	00:01:00
8	Pipeline 5	XGB Classifier	**0.673**	*None*	00:00:01

Experiment Builder

Build Model



10's-100's experiments

designs

neural network

varies

batch_size
dropout
learning_rate
…
hyperparameters

trains

100's-1000's models

evaluates

Experiment Builder

supports the end-to-end workflow

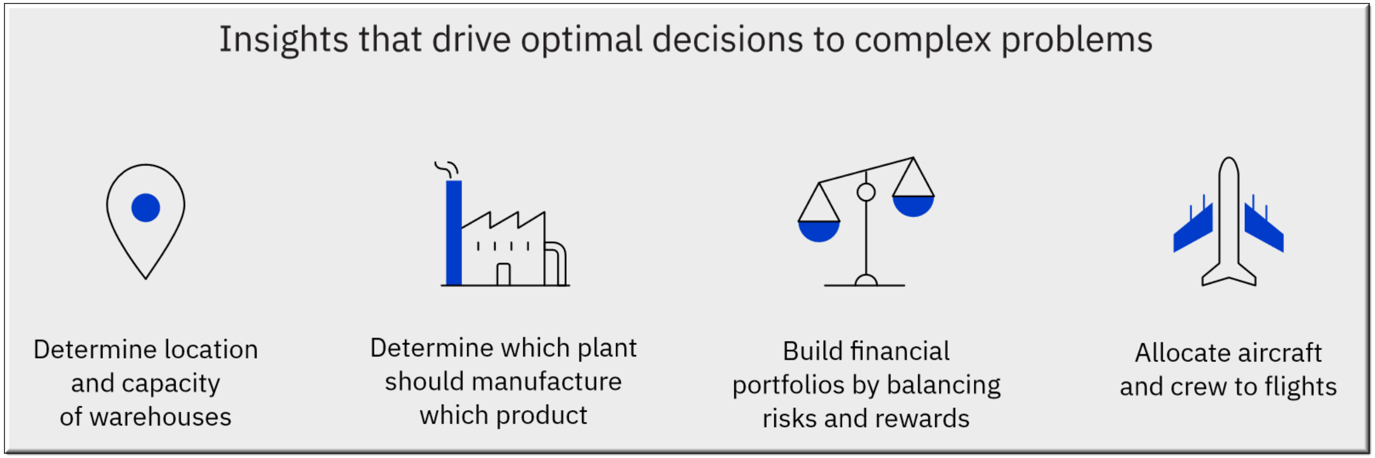
Decision Optimization

Build Model

Decision Optimization (DO) enables data science teams to capitalize on the power of *prescriptive analytics* and build solutions using a combination of techniques like optimization and machine learning.

Integrated with Watson Studio, Decision Optimization can combine optimization techniques with coding and non-coding tools, model management and deployment – as well as other data science capabilities.

Decision Optimization evaluates millions of possibilities – balancing trade-offs and business constraints to find the best possible solution.



Insights that drive optimal decisions to complex problems

Determine location and capacity of warehouses

Determine which plant should manufacture which product

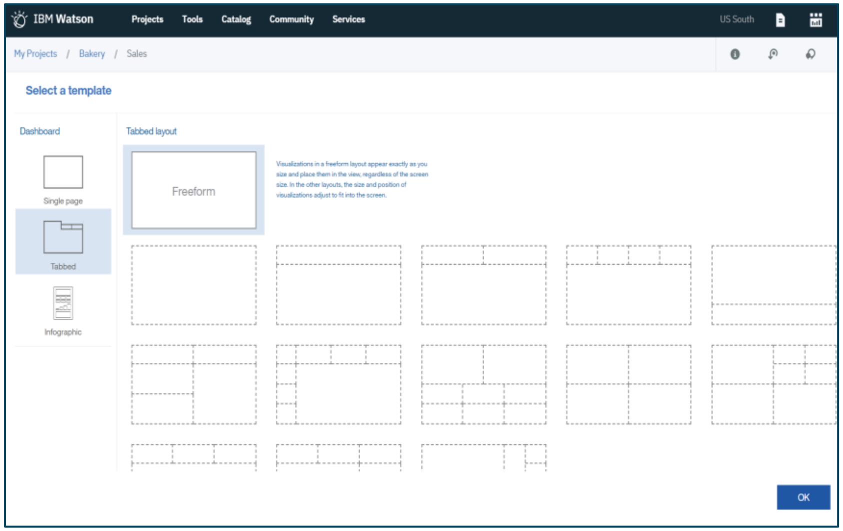
Build financial portfolios by balancing risks and rewards

Allocate aircraft and crew to flights

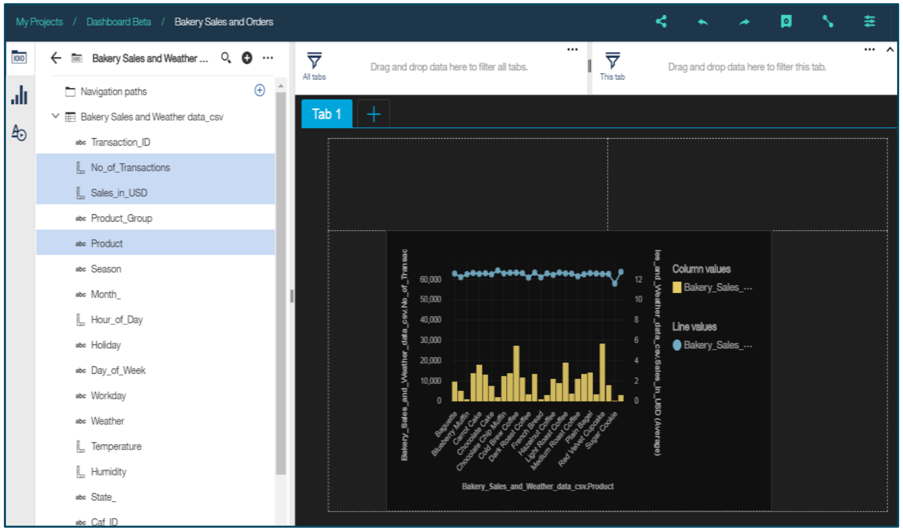
Watson Studio Dynamic Dashboards

Making insights available to all

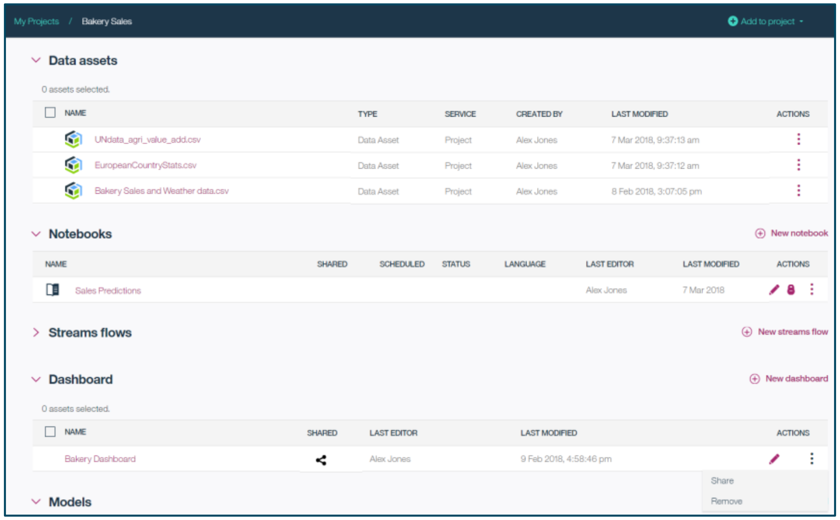
Build Model



Visualizations in a freeform layout appear exactly as you size and place them in the view, regardless of the screen size. In the other layouts, the size and position of visualizations adjust to fit into the screen.



My Projects / Dashboard Beta / Bakery Sales and Orders
Bakery Sales and Weather ...
Navigation paths
Bakery Sales and Weather data_csv
Transaction_ID
No_of_Transactions
Sales_in_USD
Product_Group
Product
Season
Month_
Hour_of_Day
Holiday
Day_of_Week
Workday
Weather
Temperature
Humidity
State_
Caf_ID
All tabs
Drag and drop data here to filter all tabs.
This tab
Drag and drop data here to filter this tab.
Tab 1
Bakery_Sales_and_Weather_data_csv.No_of_Transac
Bakery_Sales_and_Weather_data_csv.Sales_in_USD (Average)
Bakery_Sales_and_Weather_data_csv.Product
60,000
50,000
40,000
30,000
20,000
10,000
0
12
10
8
6
4
2
0
Column values
Bakery_Sales_...
Line values
Bakery_Sales_...



Data assets

0 assets selected.

NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
UNdata_agri_value_add.csv	Data Asset	Project	Alex Jones	7 Mar 2018, 9:37:13 am	⋮
EuropeanCountryStats.csv	Data Asset	Project	Alex Jones	7 Mar 2018, 9:37:12 am	⋮
Bakery Sales and Weather data.csv	Data Asset	Project	Alex Jones	8 Feb 2018, 3:07:05 pm	⋮

Notebooks

⊕ New notebook

NAME	SHARED	SCHEDULED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
Sales Predictions					Alex Jones	7 Mar 2018	⋮

Streams flows

⊕ New streams flow

Dashboard

⊕ New dashboard

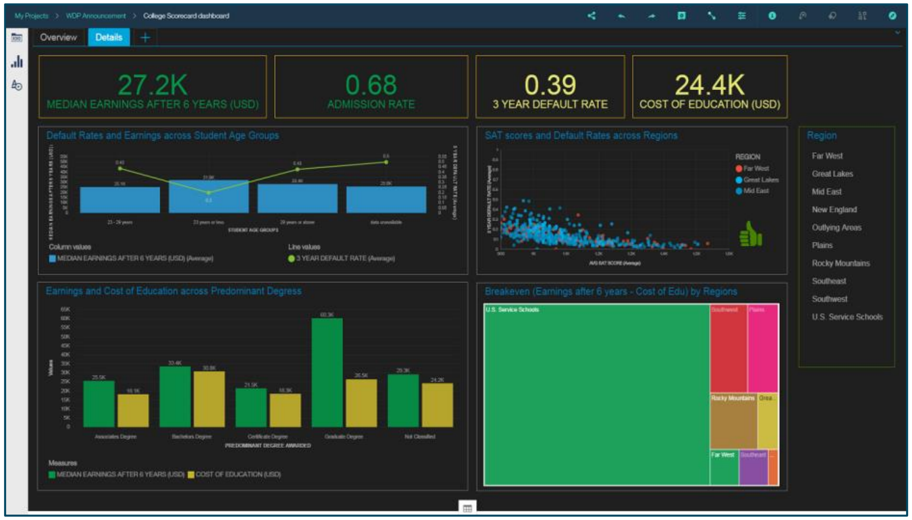
0 assets selected.

NAME	SHARED	LAST EDITOR	LAST MODIFIED	ACTIONS
Bakery Dashboard		Alex Jones	9 Feb 2018, 4:58:46 pm	⋮

Share

Remove

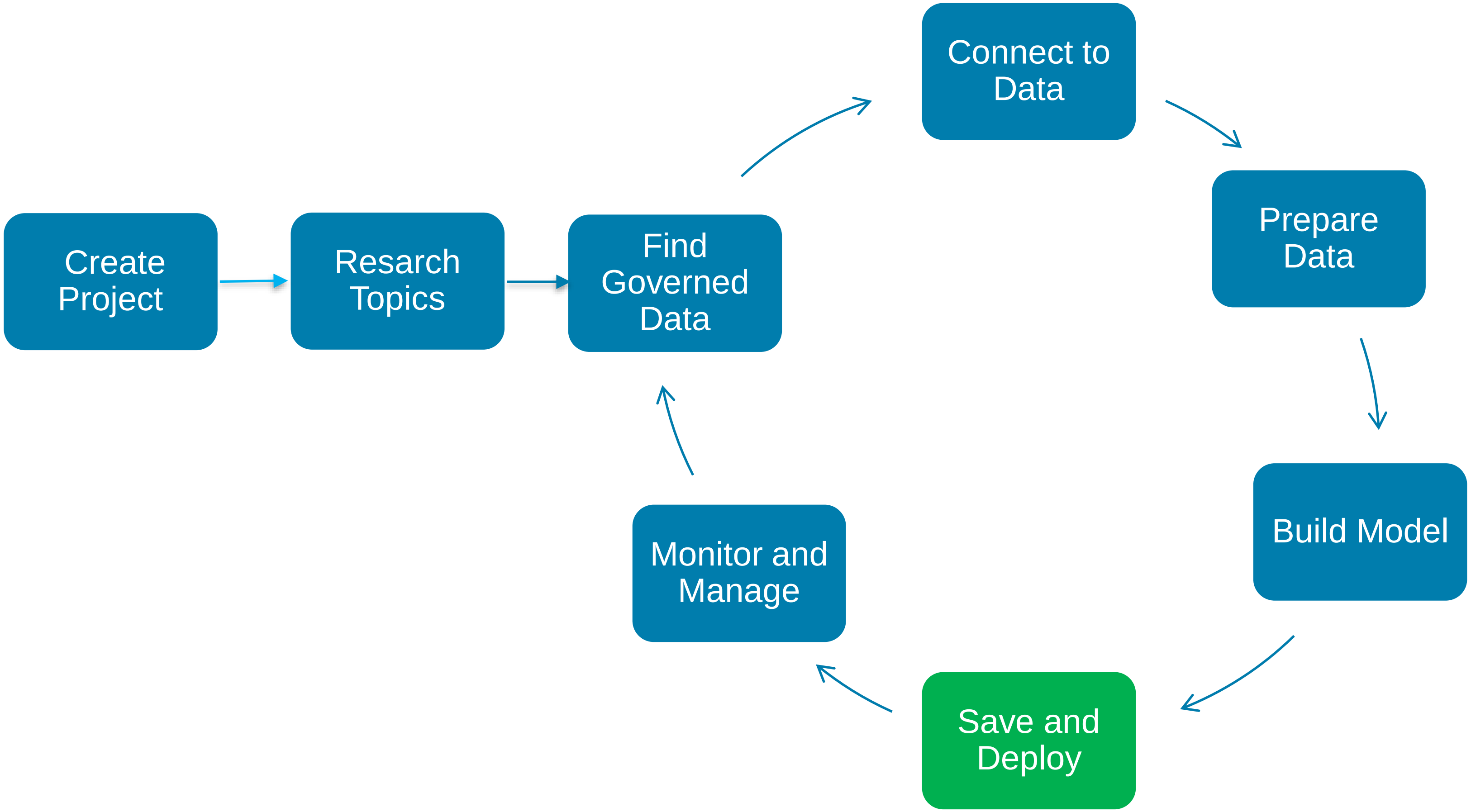
Models



My Projects > WDP Announcement > College Scorecard dashboard
Overview
Details
27.2K
MEDIAN EARNINGS AFTER 6 YEARS (USD)
0.68
ADMISSION RATE
0.39
3 YEAR DEFAULT RATE
24.4K
COST OF EDUCATION (USD)
Default Rates and Earnings across Student Age Groups
Column values
MEDIAN EARNINGS AFTER 6 YEARS (USD) (Average)
Line values
3 YEAR DEFAULT RATE (Average)
SAT scores and Default Rates across Regions
REGION
Far West
Great Lakes
Mid East
Region
Far West
Great Lakes
Mid East
New England
Outlying Areas
Plains
Rocky Mountains
Southeast
Southwest
U.S. Service Schools
Earnings and Cost of Education across Predominant Degrees
Associates Degree
Bachelors Degree
Certificate Degree
Graduate Degree
Not Classified
PREDOMINANT DEGREE AWARDED
Measures
MEDIAN EARNINGS AFTER 6 YEARS (USD)
COST OF EDUCATION (USD)
Breakeven (Earnings after 6 years - Cost of Edu) by Regions
U.S. Service Schools
Southeast
Plains
Rocky Mountains
Far West
Southwest

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Create Project

Resarch Topics

Find Governed Data

Connect to Data

Prepare Data

Build Model

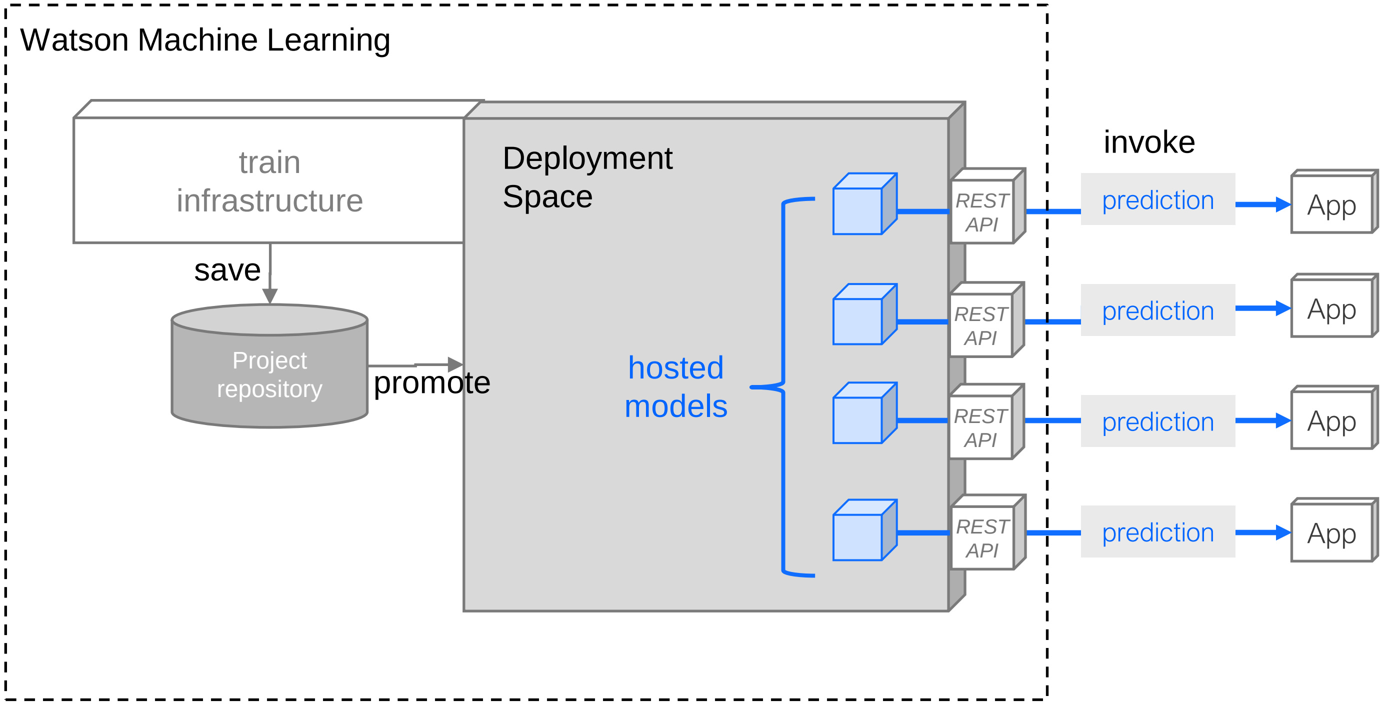
Save and Deploy

Monitor and Manage

Save and Deploy Trained Models

Save and Deploy Models with Watson Machine Learning

Save and Deploy



Watson Studio Save and Deploy Features

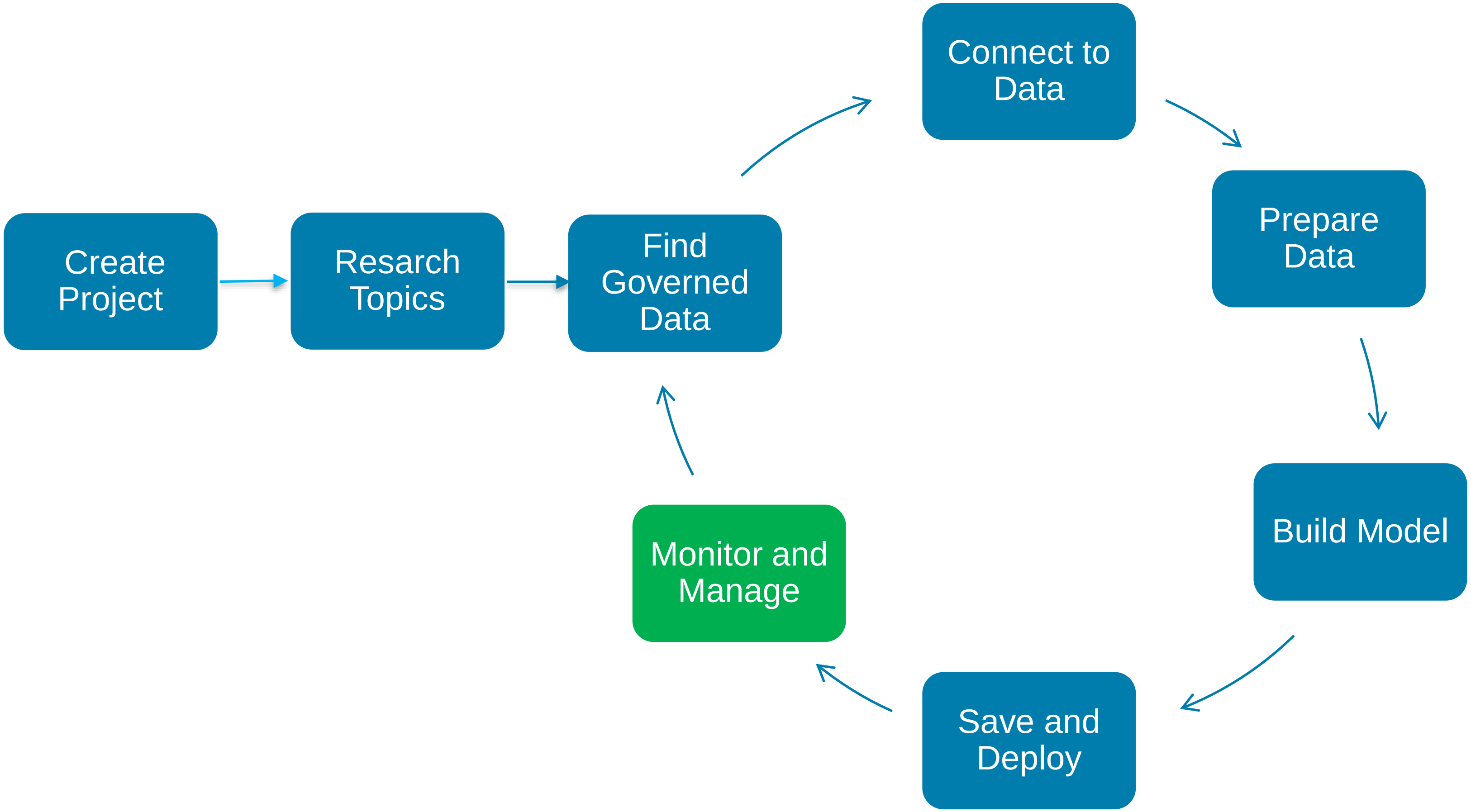
Save and Deploy

Save and Deploy Models with Watson Machine Learning

- Watson Machine Learning API to save/load models to/from repository
- Watson Machine Learning API to deploy saved models easily and have them scale automatically.
- Watson Machine Learning API to invoke deployed models

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Create Project

Resarch Topics

Find Governed Data

Connect to Data

Prepare Data

Build Model

Save and Deploy

Monitor and Manage

Monitor and Manage

Our vision for Trusted AI

Pillars of trust, woven into the lifecycle of an AI application

Is it accurate?

Is it fair?

Is it easy to understand ?

Did anyone tamper with it?

Watson OpenScale

Monitor and Manage

Trust and Transparency

- Intelligently delivers bias mitigation help
- Provides traceability & auditability of AI predictions made in production applications
- Tracks AI accuracy in applications
- Explains an outcome in business terms
- Drift Detection

Automation

- Automatically detects and mitigates bias in model output, without affecting currently deployed model or outcomes

Open By Design

- Monitor models deployed on third party model server engines
- Deploy behind enterprise firewall or on IaaS provider

Outline

- Data Science Overview
- Cloud Pak for Data Overview
- Lab Overview

Lab Use Case: Female Human Trafficking

Input

- Generated fake travel records based on incoming custom forms.
- Subset of records were vetted as “high”, “medium”, or “low” risk for Female Human Trafficking by an analyst.

Goal is to train a model on the vetted data to be able to score the unvetted travel records into high, medium, or low categories.

Lab Data

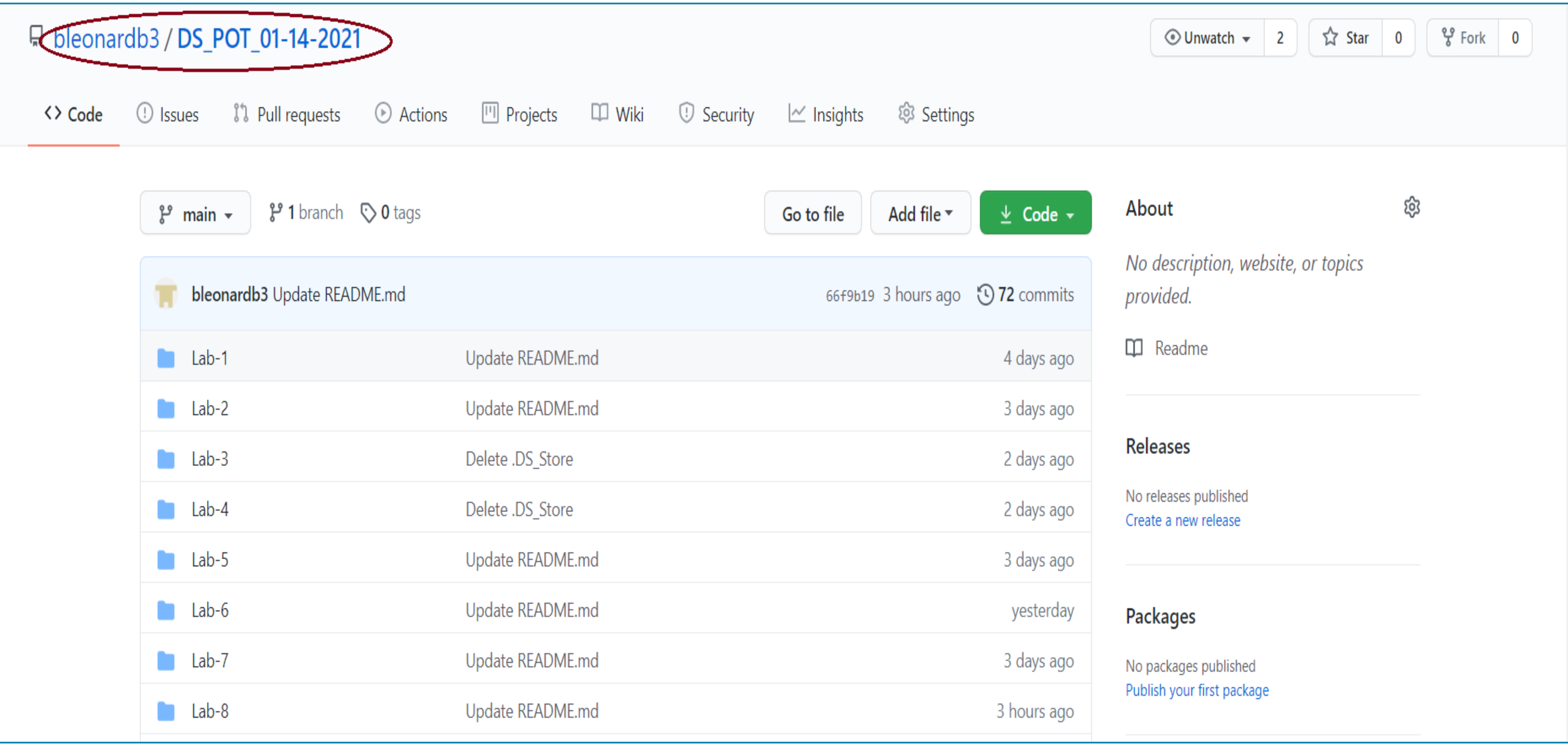
Field	Description
UUID	Hash-based unique identifier
VETTING_LEVEL	Analyst vetting status : 100- PENDING, 10 – HIGH, 20 – MED, 10 - LOW
NAME	Person name
GENDER	Person Gender
AGE (SPSS Modeler)	Person age at time of travel
BIRTH_DATE (Notebook)	Person birth date
BIRTH_COUNTRY	Person full birth country
BIRTH_COUNTRY_CODE	Person ISO 2 country
OCCUPATION CATEGORY	Person occupation as declared on form
ADDRESS	Person US address
SSN	Person Social Security Number
PASSPORT_NUMBER	Person Passport Number
PASSPORT_COUNTRY	Person Passport Issuing Country
PASSPORT_COUNTRY_CODE	Person Passport Issuing Country ISO 2 Code
COUNTRYIES_VISITED	The countries visited as declared on form
COUNTRIES_VISITED_COUNT	The number of countries visited as declared on form
ARRIVAL_AIRPORT_COUNTRY_CODE	ARRIVAL Airport country code ISO2
AIRPORT_ARRIVAL_IATA	ARRIVAL Airport 3 character code
AIRPORT_ARRIVAL_MUNICIPALITY	ARRIVAL Airport Municipality Derived from Code
ARRIVAL_AIRPORT_REGION	ARRIVAL Airport Region Derived from Code
DEPARTURE_AIRPORT_COUNTRY_CODE	DEPARTURE Airport Country code ISO2
DEPARTURE_AIRPORT_IATA	DEPARTURE Airport 3 character code
DEPARTURE_AIRPORT_MUNICIPALITY	DEPARTURE Airport Municipality Derived from Code.

🟥	Target
🟦	Features

Lab Tips

- Cloud Pak for Data url: dataplatform.cloud.ibm.com
- Labs are in www.github.com/bleonardb3/DS_POT_01-14-2021 repository.
- Instructions for each Lab are in the README file in the respective Lab folder.
- Cloud development enables making frequent improvements in the user interface. We reviewed the lab instructions and made screen updates so they should be pretty faithful to the user interface. Small differences may occur but shouldn't get in the way of successfully completing the labs.
- Do not use Internet Explorer or Edge as the browser. For Mac users do not use Safari.
- Watson Studio -→ Cloud Pak for Data (Watson Studio is component)
- All of the Labs should be done in the project that you created in Lab-1

Github Repository



bleonardb3 / DS_POT_01-14-2021

bleonardb3 Update README.md 66f9b19 3 hours ago 72 commits

Lab-1	Update README.md	4 days ago
Lab-2	Update README.md	3 days ago
Lab-3	Delete .DS_Store	2 days ago
Lab-4	Delete .DS_Store	2 days ago
Lab-5	Update README.md	3 days ago
Lab-6	Update README.md	yesterday
Lab-7	Update README.md	3 days ago
Lab-8	Update README.md	3 hours ago

About

No description, website, or topics provided.

Readme

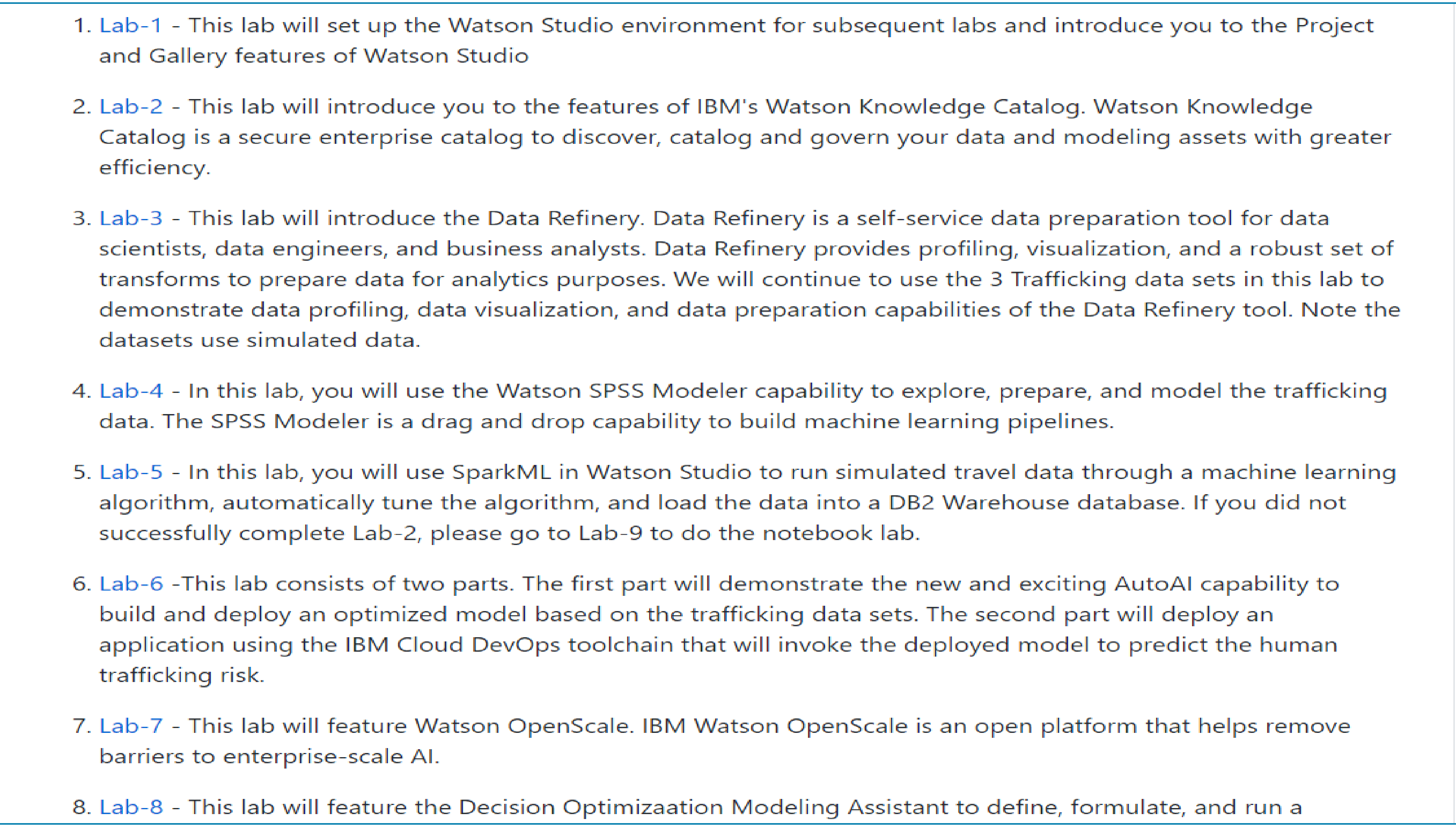
Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

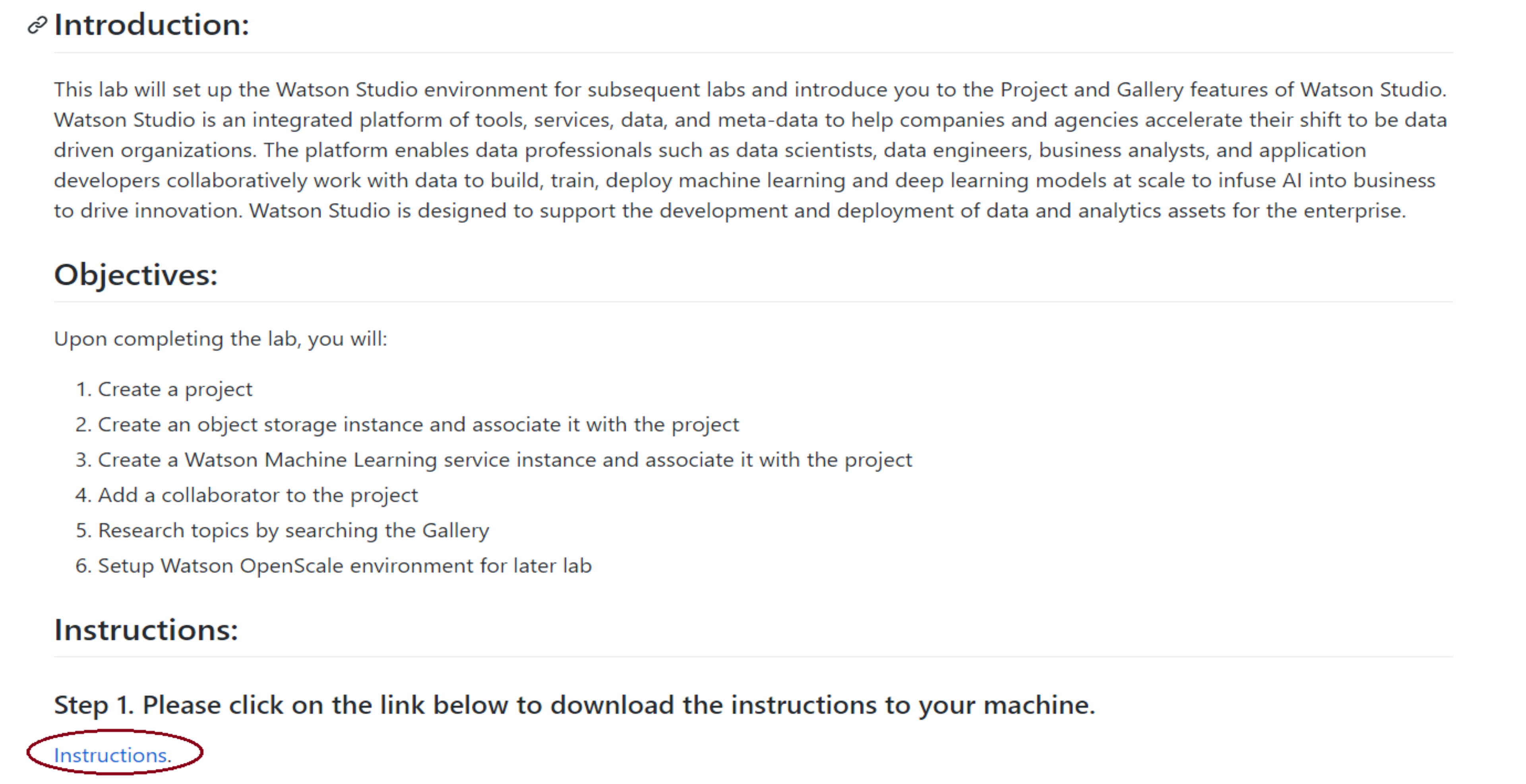
Github Repository
Readme



1. Lab-1 - This lab will set up the Watson Studio environment for subsequent labs and introduce you to the Project and Gallery features of Watson Studio
2. Lab-2 - This lab will introduce you to the features of IBM's Watson Knowledge Catalog. Watson Knowledge Catalog is a secure enterprise catalog to discover, catalog and govern your data and modeling assets with greater efficiency.
3. Lab-3 - This lab will introduce the Data Refinery. Data Refinery is a self-service data preparation tool for data scientists, data engineers, and business analysts. Data Refinery provides profiling, visualization, and a robust set of transforms to prepare data for analytics purposes. We will continue to use the 3 Trafficking data sets in this lab to demonstrate data profiling, data visualization, and data preparation capabilities of the Data Refinery tool. Note the datasets use simulated data.
4. Lab-4 - In this lab, you will use the Watson SPSS Modeler capability to explore, prepare, and model the trafficking data. The SPSS Modeler is a drag and drop capability to build machine learning pipelines.
5. Lab-5 - In this lab, you will use SparkML in Watson Studio to run simulated travel data through a machine learning algorithm, automatically tune the algorithm, and load the data into a DB2 Warehouse database. If you did not successfully complete Lab-2, please go to Lab-9 to do the notebook lab.
6. Lab-6 -This lab consists of two parts. The first part will demonstrate the new and exciting AutoAI capability to build and deploy an optimized model based on the trafficking data sets. The second part will deploy an application using the IBM Cloud DevOps toolchain that will invoke the deployed model to predict the human trafficking risk.
7. Lab-7 - This lab will feature Watson OpenScale. IBM Watson OpenScale is an open platform that helps remove barriers to enterprise-scale AI.
8. Lab-8 - This lab will feature the Decision Optimizaation Modeling Assistant to define, formulate, and run a

Github Repository

Lab-1 Readme



Introduction:

This lab will set up the Watson Studio environment for subsequent labs and introduce you to the Project and Gallery features of Watson Studio. Watson Studio is an integrated platform of tools, services, data, and meta-data to help companies and agencies accelerate their shift to be data driven organizations. The platform enables data professionals such as data scientists, data engineers, business analysts, and application developers collaboratively work with data to build, train, deploy machine learning and deep learning models at scale to infuse AI into business to drive innovation. Watson Studio is designed to support the development and deployment of data and analytics assets for the enterprise.

Objectives:

Upon completing the lab, you will:

1. Create a project
2. Create an object storage instance and associate it with the project
3. Create a Watson Machine Learning service instance and associate it with the project
4. Add a collaborator to the project
5. Research topics by searching the Gallery
6. Setup Watson OpenScale environment for later lab

Instructions:

Step 1. Please click on the link below to download the instructions to your machine.

Instructions.

Lab-1: Set up Environment

Introduction:

This lab will set up the Watson Studio environment for subsequent labs and introduce you to the Project and Community features of Watson Studio.

Objectives:

Upon completing this lab, you will know how to:

- Create a project
- Create an object storage instance and associate it with the project
- Create a Watson Machine Learning service instance and associate it with the project
- Add a collaborator to the project
- Research topics by searching the Gallery
- Setup Watson OpenScale environment for a later lab.

Lab-2: Introduction to Watson Knowledge Catalog

Introduction:

This lab will introduce you to the features of IBM's Watson Knowledge Catalog. Watson Knowledge Catalog is a secure enterprise catalog to discover, catalog and govern your data and modeling assets with greater efficiency.

Objectives:

The goal of the lab is to gain familiarity with the features of the Watson Knowledge Catalog. Upon completing the lab, you will know how to:

- Create a governed catalog
- Add a member to the catalog
- Add Data Assets to the catalog
- Search the catalog
- Edit/Review/Profile a Data Asset
- Demonstrate access control features
- Create and enforce policy
- Push the Data Assets to a project.

End-to-End Data Science using IBM’s Cloud Pak for Data

We will return for review at 11:45 am.

Lab-3: Introduction to the Data Refinery

Introduction:

In this lab, you will use the Watson Studio Data Refinery to profile data, visualize data, and prepare data for modeling.

Objectives:

Upon completing the lab, you will know how to:

- Profile the data
- Visualize the data to gain a better understanding
- Prepare the data for modeling
- Run the sequence of data preparation operations on the entire data set.

Lab-4: SPSS Modeler

Introduction:

In this lab, you will use the Watson Studio SPSS Modeler capability to explore, prepare, and model trafficking data. The SPSS Modeler is a drag and drop capability to build machine learning pipelines.

Objectives:

Upon completing the lab, you will:

- Become familiar with the Watson Studio SPSS Modeler capability
- Profile the data set
- Explore the data set with visualizations
- Transform the data
- Train/Evaluate a machine learning mode.

End-to-End Data Science using IBM’s Cloud Pak for Data

We will return for lecture at
2:00 pm. Please work on labs 3 and 4

Categories of Machine Learning

Supervised learning

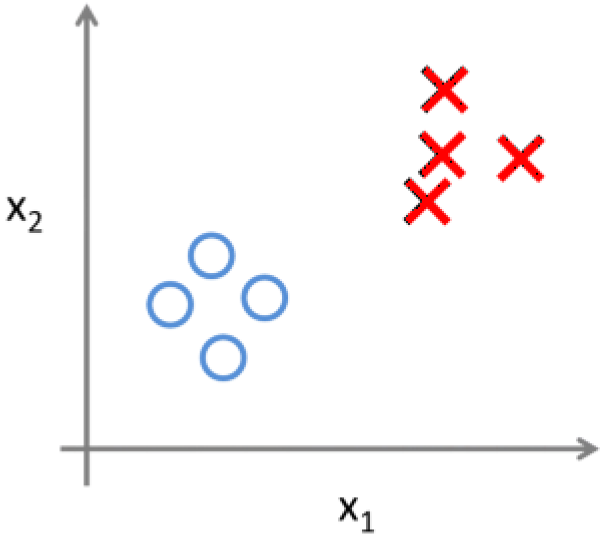
- The program is “trained” on a pre-defined set of “training examples”, which then facilitate its ability to reach an accurate conclusion when given new data
- The algorithm is presented with example inputs and their outcomes (labels)
- The goal is to learn a general rule that maps inputs to outputs

Unsupervised learning

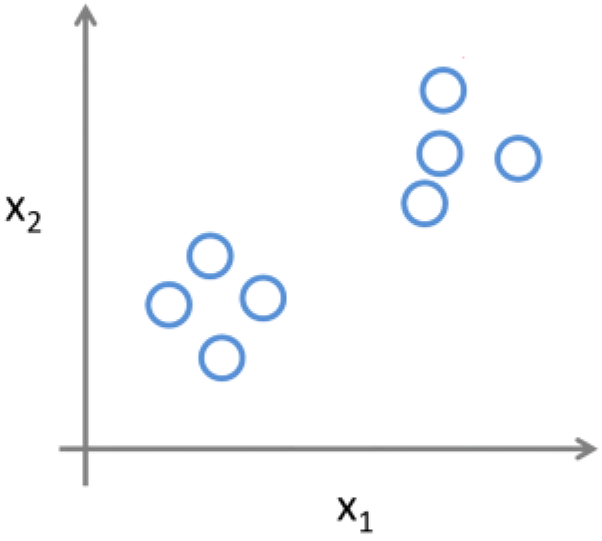
- No labels are given to the learning algorithm, leaving it on its own to find structure (patterns and relationships) in its input

Supervised vs. Unsupervised Learning

Supervised Learning

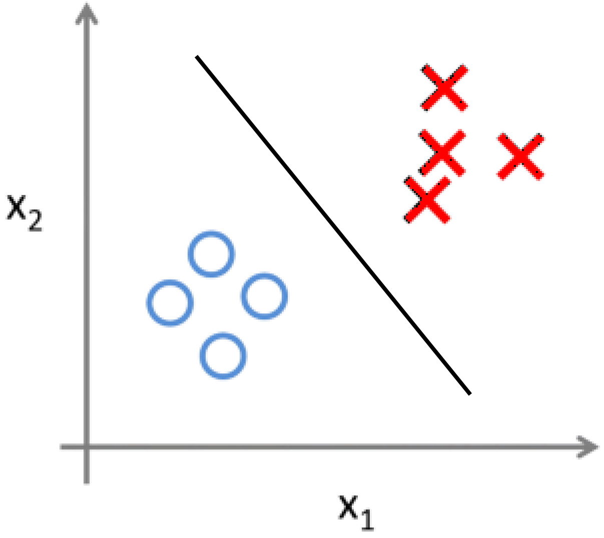


Unsupervised Learning

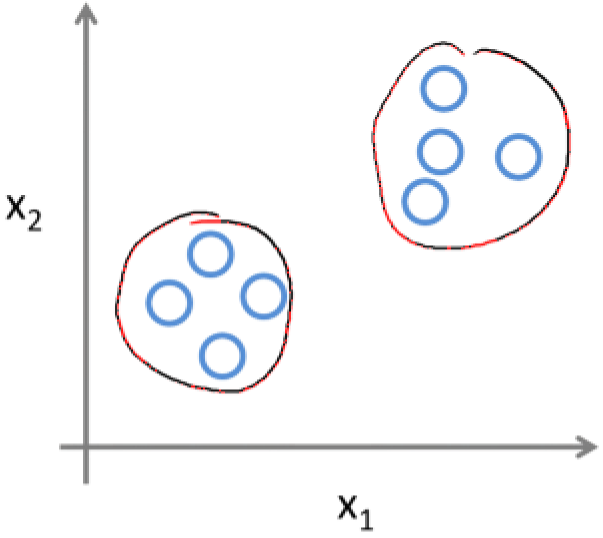


Supervised vs. Unsupervised Learning

Supervised Learning



Unsupervised Learning



Preprocessing: Matrix for Machine Learning

Known as:

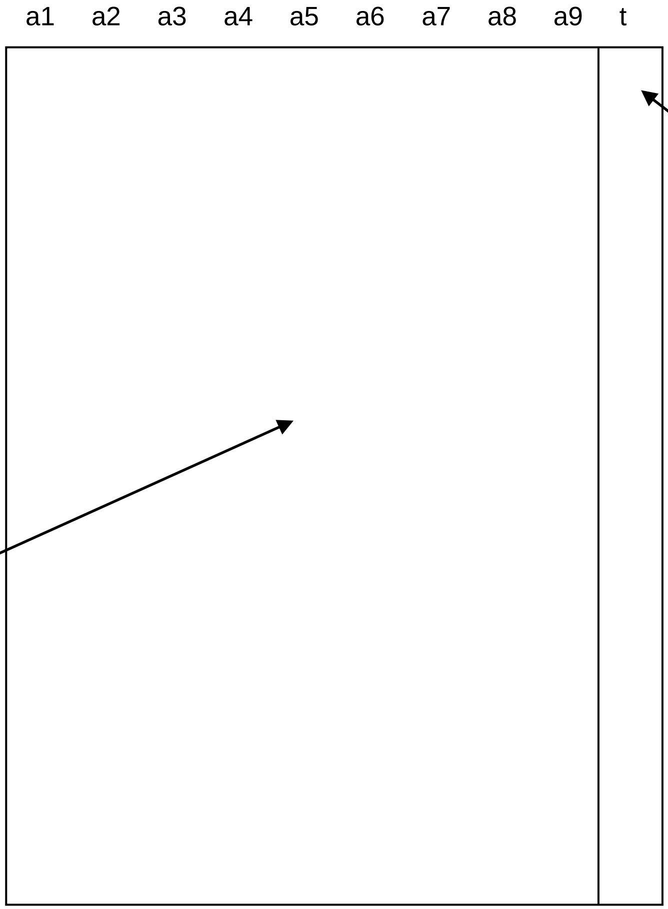
- Attributes
- Features
- Predictor variables
- Explanatory variables

Scale variables:

- Continuous variables, which can be measured on an interval scale or ratio scale
- ‘Weight’, ‘Temperature’, ‘Salary’, etc…

Categorical variables:

- Data with a limited number of distinct values or categories (nominal or ordinal)
- ‘Hair color’, ‘Gender’, ‘Grape varieties’, etc…



Known as:

- Label
- Target variable
- Dependent variable

Scale or Categorical

Training, testing, & validation sets

During the model development process, supervised learning techniques employ training and testing sets and sometimes a validation set.

- Historical data with known outcome
- Data is randomly split into training, testing, and/or validation sets (mutually exclusive records)

Why?

- Training set
  - Build the model
  - Tune the parameters
- Validation set
  - Assess model quality during training/tuning process
  - Avoid overfitting the model to the training set
- Test set
  - Estimate accuracy or error rate of model after tuning
  - Used to compare multiple models

K-Fold Cross Validation

- **Instead of using a separate validation set**

- **Shuffle Training Samples and sub-divide into “K” folds (groups)**

- **Train “K” models using K-1 folds as training data and 1 Fold as Test Data**

- **For example, K=4**
  - Model 1 Train on 1,2,3 Test on 4 – calculate and store E1 (Error)
  - Model 2 Train on 2,3,4 Test on 1 – E2
  - Model 3 Train on 3,4,1 Test on 2 - E3
  - Model 4 Train on 4,1,2 Test on 3 - E4
  - E = (E1+E2+E3+E4)/4

- **A common value for K is 10**

Spark and Spark ML

Spark – why should I use it?

- Spark is a highly scalable runtime environment for analytics
- Provides the runtime engine and API
- Supports multiple languages: Python (PySpark), R (SparkR) and Scala

If you want to take advantage of Spark scalability and performance, you have to use Spark APIs

- Example (Python): Spark data frame vs. Pandas, Spark algorithms vs. scikit-learn
- It’s possible to “mix and match” Spark and non-Spark code in a single notebook: the runtime environment will switch automatically
  - For example, use Python API for data understanding and SparkML for modeling

Spark Machine Learning API: https://spark.apache.org/docs/latest/ml-guide.html

Supported versions of Spark: https://www.ibm.com/software/reports/compatibility/clarity/prereqsForProduct.html

Lab-5: Flow

Read in data from Cataloged Assets

- Join trafficking, job categories, occupations data

Identify Labels

- Label the data (“VETTING_LEVEL”)
- Select features

Feature Engineering (Transformation)

- StringIndexer (occupation, country, gender, birth year variables)
- VectorAssembler
- Normalizer

Define Model and Setup Pipeline

- Naïve Bayes
- Random Forest

Train the Model

- Split input data into Training (70%) and Test (30%) DataFrames
- Cache the resulting DataFrames
- Fit the Pipeline to the Training data set

Lab-5: Flow (continued)

Evaluate the resulting predictions

- Area under the ROC curve

Tune the model (hyperparamaters)

- Build Parameter Grid
- Cross-evaluate to find the best model

Score the unvetted records

- Use Best Model to Score unvetted records (VETTING LEVEL == 100)

Save the model in the Model Repository

Lab-5: Machine Learning using SparkML

Introduction:

In this lab, you will use SparkML in Watson Studio to run generated travel data through a machine learning algorithm, automatically tune the algorithm, and load the prediction results into a DB2 on Cloud database.

Objectives:

Upon completing the lab, you will know how to use a Jupyter Notebook to:

- Connect to a cataloged assets to read in data used for machine learning.
- Select the target and features
- Transform data
- Declare a machine learning model.
- Setup up the data transform and modeling pipeline
- Train the model.
- Evaluate the model.
- Automatically tune the model.
- Score data
- Save the trained model

Lab-6: AutoAI

Introduction:

This lab will demonstrate the exciting AutoAI capability to build and deploy an optimized model based on the trafficking data set.

Objectives:

Upon completing the lab, you will:

- Become familiar with the AutoAI feature of Watson Studio.
- Train/Evaluate a machine learning model
- Save and Deploy a machine learning model.
- Test the Machine Learning model

End-to-End Data Science using IBM’s Cloud Pak for Data

We will return for lecture at 3:30 pm. Please work on labs 5 and 6

Our vision for Trusted AI

Is it accurate?

Is it fair?

Is it easy to understand ?

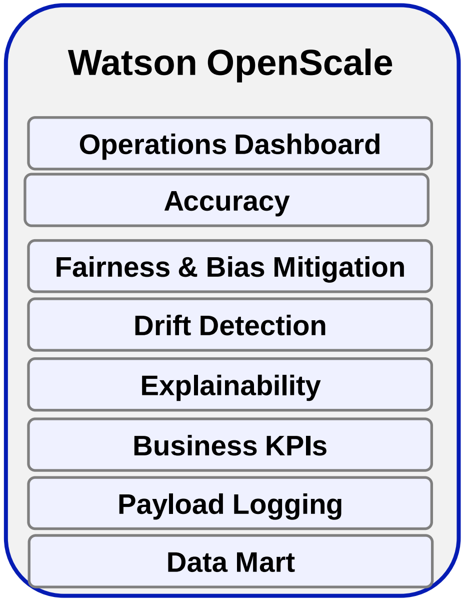
Did anyone tamper with it?

Watson OpenScale: Overview

Watson OpenScale:

- Automates and operates AI at scale across its entire lifecycle
- Delivers transparent, explainable outcomes freed from bias and drift
- Provides confidence in AI outcomes and spans the gap between the teams that operate AI and the business units that use these applications
- Monitors models developed in a 3rd party IDE, open source framework and hosted in a 3rd party or private model serve engine

Manage AI at Scale



Watson OpenScale

Operations Dashboard

Accuracy

Fairness & Bias Mitigation

Drift Detection

Explainability

Business KPIs

Payload Logging

Data Mart

Model build / train frameworks

TensorFlow

scikit learn

Keras

PYTORCH

Caffe2

Spark MLlib

Model serving environments

kubernetes

Azure ML

aws

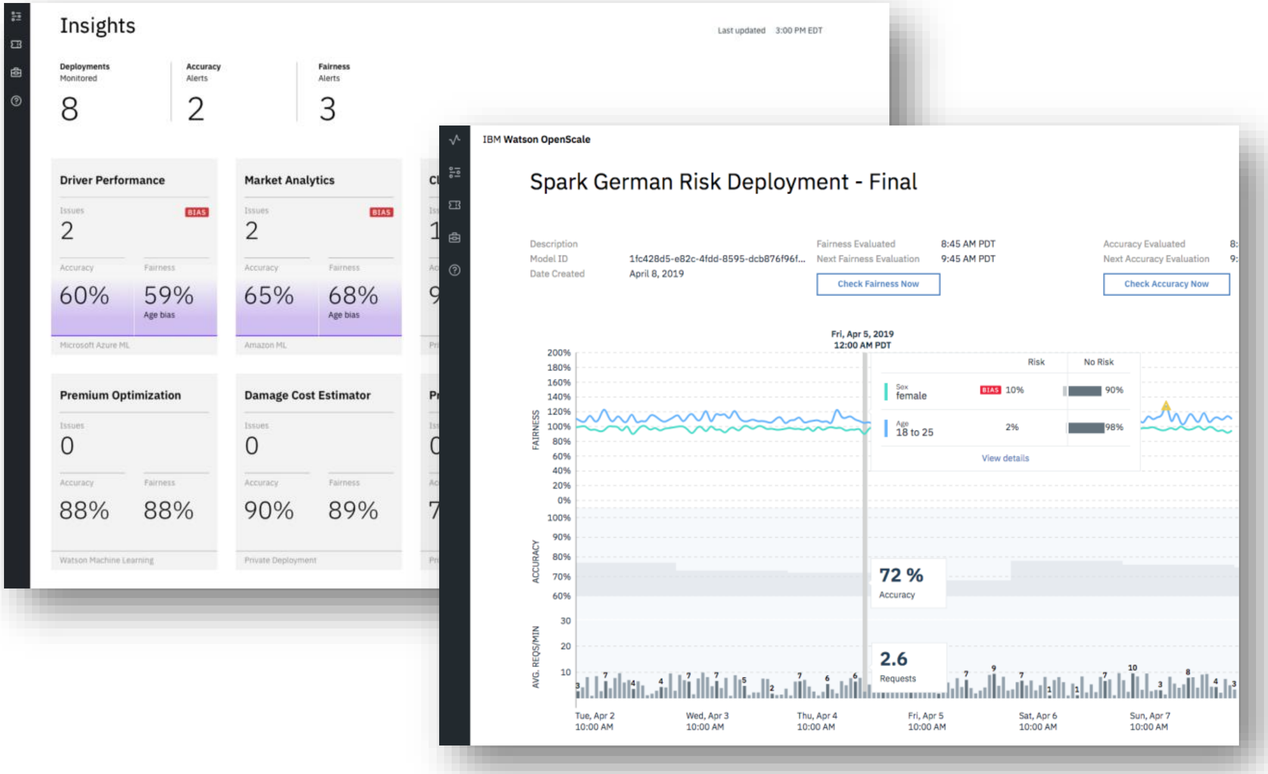
Watson OpenScale: Operations Dashboard

Description:

Monitor deployed models in a single dashboard that can be filtered by deployment making it easy to manage AI in apps

Value:

- Configure alerts or actions to be triggered when KPIs exceed threshold, ensuring model quality for improve business outcomes
- Measure model accuracy as it pertains to it's ability to deliver outcomes more accurate than knowledge workers
- Provides "continuous evolution" for your models



Insights

Last updated 3:00 PM EDT

Deployments Monitored: 8
Accuracy Alerts: 2
Fairness Alerts: 3

Deployment	Issues	Accuracy	Fairness	Platform
Driver Performance	2 BIAS	60%	59% Age bias	Microsoft Azure ML
Market Analytics	2 BIAS	65%	68% Age bias	Amazon ML
Premium Optimization	0	88%	88%	Watson Machine Learning
Damage Cost Estimator	0	90%	89%	Private Deployment

IBM Watson OpenScale

Spark German Risk Deployment - Final

Description
Model ID: 1fc428d5-e82c-4fdd-8595-dcb876f96f...
Date Created: April 8, 2019

Fairness Evaluated: 8:45 AM PDT
Next Fairness Evaluation: 9:45 AM PDT

Accuracy Evaluated: 8:
Next Accuracy Evaluation: 9:

Fri, Apr 5, 2019
12:00 AM PDT

	Risk	No Risk
Sex female	BIAS 10%	90%
Age 18 to 25	2%	98%

72 % Accuracy

2.6 Requests

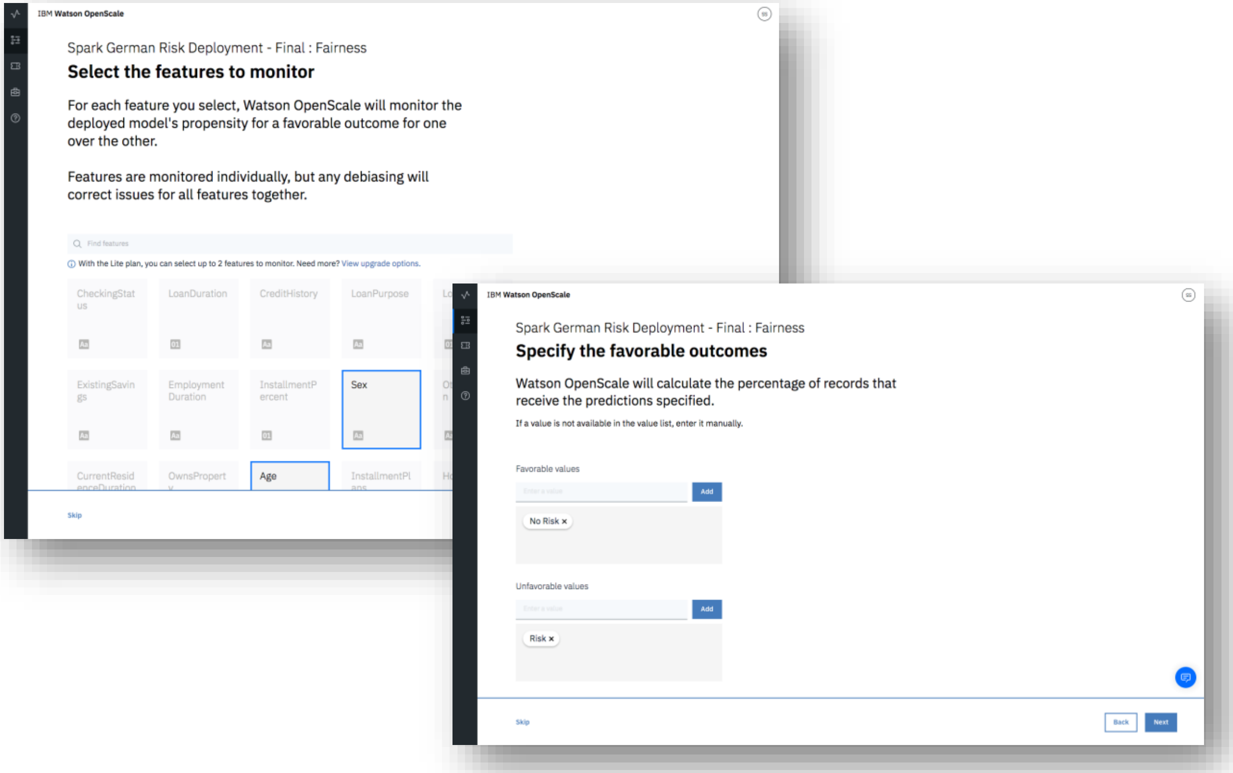
Watson OpenScale: Model Fairness

Description:

Production Models need to make fair decisions and *can not be biased* in their recommendations

How it works:

- Outcomes are selected as “favorable or unfavorable”
- “Favored Populations” and “protected populations” are selected where majority and minority groups are found
- A score is calculated based on the probability of favorable outcome for minority vs. probability of favorable outcome for majority



IBM Watson OpenScale

Spark German Risk Deployment - Final : Fairness

Select the features to monitor

For each feature you select, Watson OpenScale will monitor the deployed model's propensity for a favorable outcome for one over the other.

Features are monitored individually, but any debiasing will correct issues for all features together.

ⓘ With the Lite plan, you can select up to 2 features to monitor. Need more? View upgrade options.

Selected features: Sex, Age

IBM Watson OpenScale

Spark German Risk Deployment - Final : Fairness

Specify the favorable outcomes

Watson OpenScale will calculate the percentage of records that receive the predictions specified.

If a value is not available in the value list, enter it manually.

Favorable values: No Risk

Unfavorable values: Risk

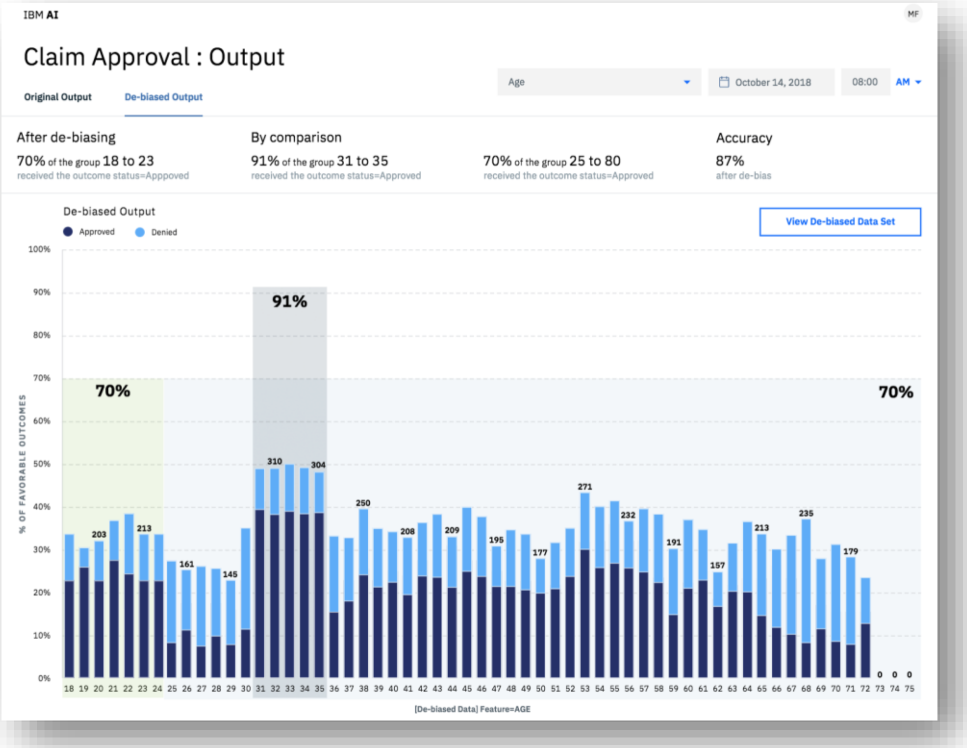
Watson OpenScale: Bias Mitigation

Description:

Fairness is enforced with automatic bias mitigation.

How it works:

- Calculated on an *hourly basis* (over a sliding window defined by the user)
- Optimizations identify the *right subset of data to perturb* (rather than perturbing all the data)
- *Perturbed data is sent to the deployed* model to determine effect of perturbations
- An internal bias detection model (logistic regression) is built using perturbed data that *classifies whether new prediction will be biased or not*
- Users receive both the *original prediction* plus the *internal model's classification* of whether the monitored model's prediction is biased or not



IBM AI

Claim Approval : Output

Original Output | De-biased Output

Age | October 14, 2018 | 08:00 AM

After de-biasing

70% of the group 18 to 23
received the outcome status=Apppoved

By comparison

91% of the group 31 to 35
received the outcome status=Approved

70% of the group 25 to 80
received the outcome status=Approved

Accuracy

87%
after de-bias

De-biased Output

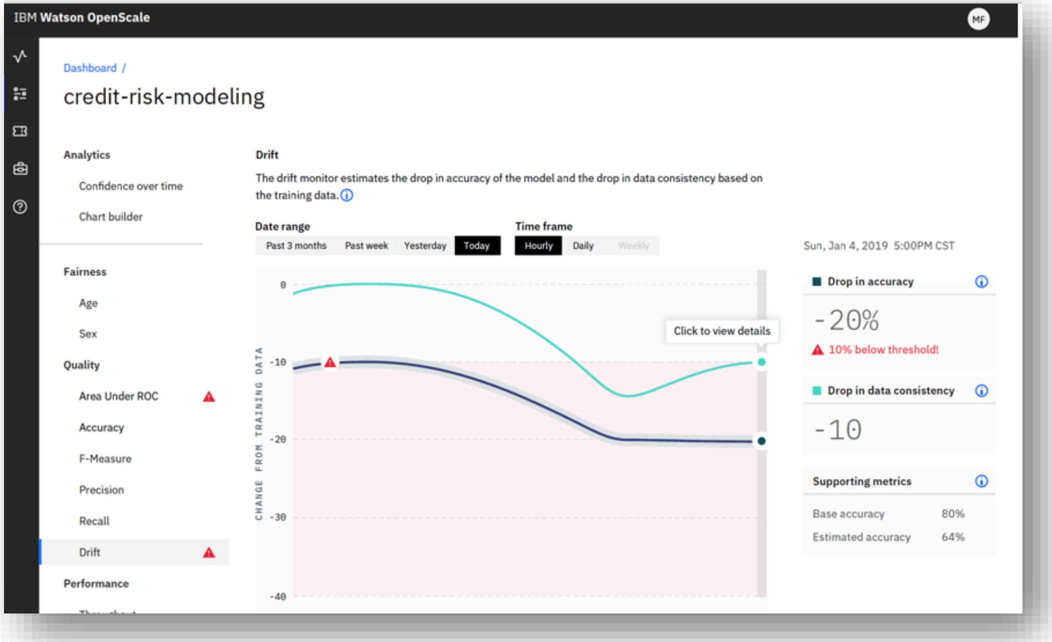
View De-biased Data Set

Watson OpenScale: Drift detection

Description:

OpenScale monitors for two types of drift:

- **Drop in accuracy**: It estimates the drop in accuracy of the model at runtime. Accuracy could drop if there is an increase in transactions similar to those which the model was unable to evaluate correctly with the training data.
- **Drop in data consistency**: It estimates the drop in consistency of the data at runtime as compared to the characteristics of the data at training time.



The drift monitor estimates the drop in accuracy of the model and the drop in data consistency based on the training data.

Sun, Jan 4, 2019 5:00PM CST

Drop in accuracy

-20%

10% below threshold!

Drop in data consistency

-10

Supporting metrics

Base accuracy	80%
Estimated accuracy	64%

OpenScale does drift detection on the entire payload data.

OpenScale measures the drift without requiring labeled data. Accuracy computation using labeled data can be expensive and might not be comprehensive

Watson OpenScale: Explainability

Description:

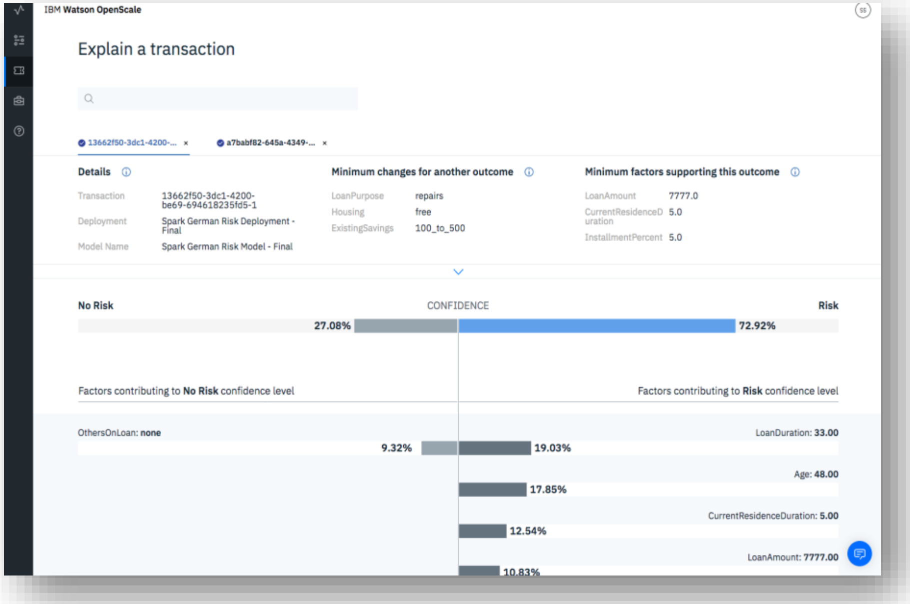
Allows you to understand which feature values of a model that are most influencing a prediction for a specific transaction

Example:

A loan is not approved by a model prediction - explainability will tell you why

How it works:

- Perturbation analysis on thousands of variations
- Risk model is created for two variations:
  - **LIME (local) Explanation:** set of features which played a positive or negative role in the prediction - also identifies the feature weights which helps to identify the most or least important features
  - **Contrastive Explanation:** Explains the behavior of the model in the vicinity of the data point whose explanation is being generated – assumption: the most common value is the least interesting from an explanation point of view



IBM Watson OpenScale

Explain a transaction

13662f50-3dc1-4200-... × a7babf82-645a-4349-... ×

Details

Transaction	13662f50-3dc1-4200-be69-694618235fd5-1
Deployment	Spark German Risk Deployment - Final
Model Name	Spark German Risk Model - Final

Minimum changes for another outcome

LoanPurpose	repairs
Housing	free
ExistingSavings	100_to_500

Minimum factors supporting this outcome

LoanAmount	7777.0
CurrentResidenceDuration	5.0
InstallmentPercent	5.0

Lab 7: Watson OpenScale

Introduction:

IBM Watson OpenScale is an open platform that helps remove barriers to enterprise-scale AI. In this lab you will configure Watson OpenScale to monitor quality, fairness,and drift and to provide the factors that explain a deployed model's classification.

Objectives:

Upon completing the lab, you will

- Provision Watson OpenScale (should be completed in Lab-1)
- View Fairness and Quality Metrics
- View Drift Metrics.
- Explain a Transaction.
- Compare Pre-production Models
- Generate a Report.

Lab 8: Watson Decision Optimization

Introduction:

This lab is based on the house construction scheduling problem tutorial in the Cloud Pak for Data documentation. The lab guides you to use the Decision Optimization Modeling Assistant to define, formulate and run a model for a house construction scheduling problem.

Objectives:

Upon completing the lab, you will have

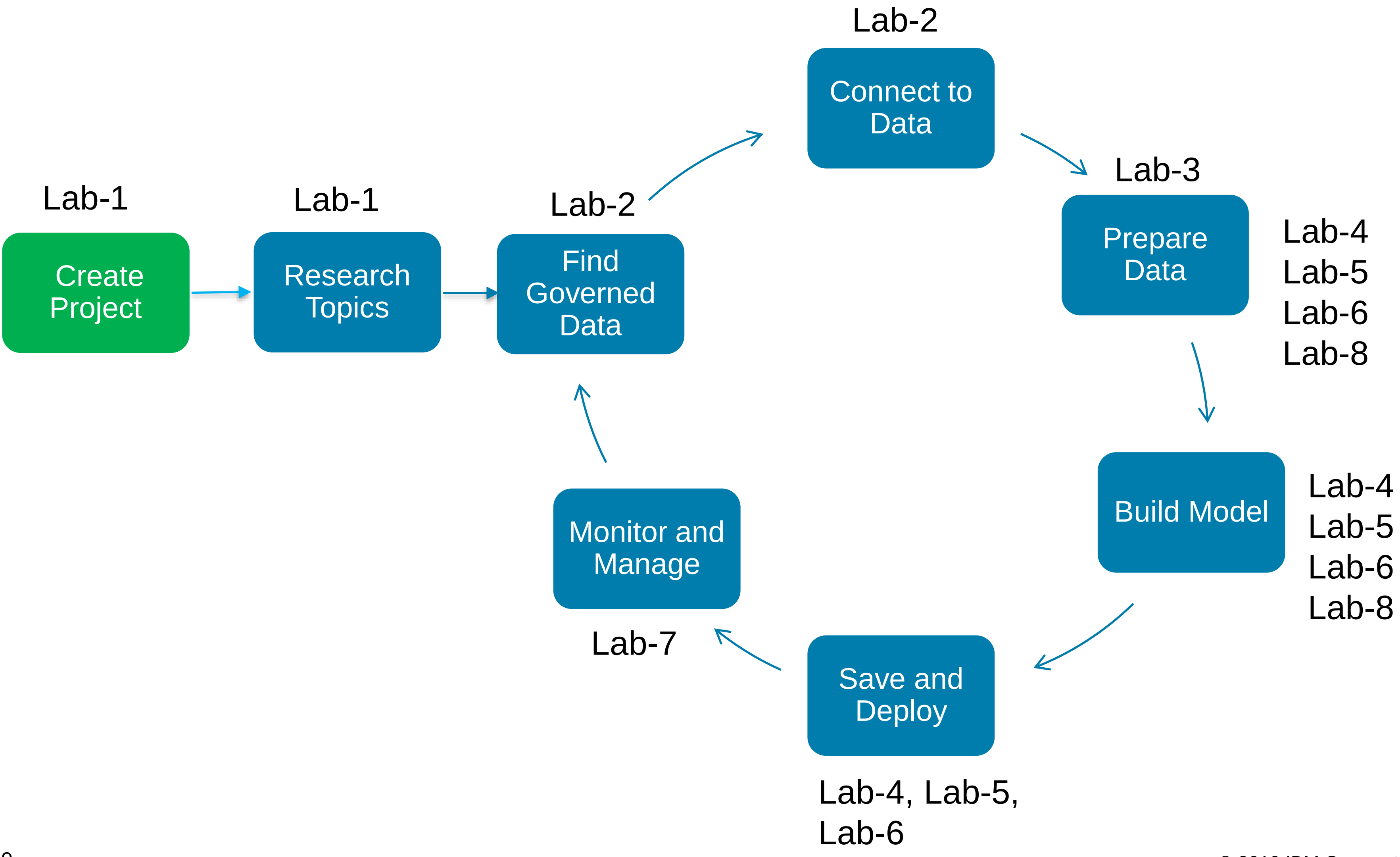
- Downloaded the sample data files
- Uploaded the files to the Watson Studio project
- Created a Decision Optimization experiment
- Prepared the Data
- Formulated and run 3 Optimization Scenarios

End-to-End Data Science using IBM's Watson Studio

We will return for lecture at 5:00 pm. Please work on labs 7 and 8

Cloud Pak for Data supports the Data Science Lifecycle

Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.



Lab-1
Create Project

Lab-1
Research Topics

Lab-2
Find Governed Data

Lab-2
Connect to Data

Lab-3
Prepare Data
Lab-4
Lab-5
Lab-6
Lab-8

Build Model
Lab-4
Lab-5
Lab-6
Lab-8

Save and Deploy
Lab-4, Lab-5,
Lab-6

Monitor and Manage
Lab-7