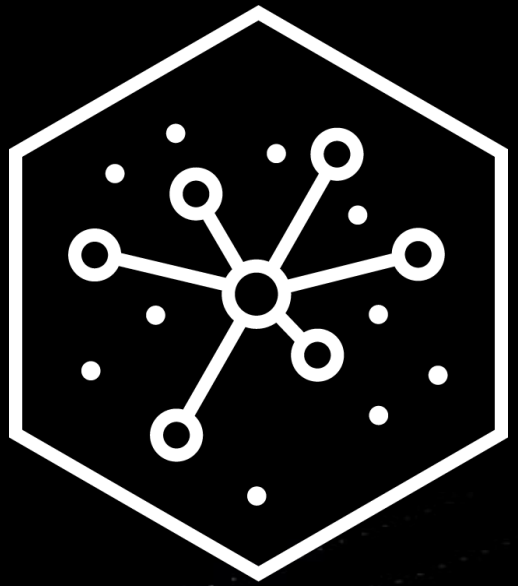
End-to-End Data Science using IBM's Cloud Pak for Data

Starts at 9:00am EST

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

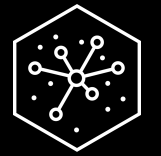


Starts at 9:00am EST

Agenda

Time	Description
09:00 AM – 09:45 AM	**Overview of Cloud Pak for Data** **Lab 1,2 Presentation**
09:45 AM – 11:30 AM	**Lab-1: Set up Environment** **Lab-2: Watson Knowledge Catalog**
11:30 AM – 11:45 AM	**Lab 3,4 Presentation**
11:45 AM – 01:45 PM	**Lab-3: Data Refinery** **Lunch** **Lab-4: SPSS Modeler**
01:45 PM – 02:00 PM	**Lab 5,6 Presentation**
02:00 PM – 03:00 PM	**Lab-5: Machine Learning with SparkML** **Lab 6: Auto AI**
03:00 PM – 03:15 PM	**Lab 7 Presentation**
03:15 PM – 04:15 PM	**Lab-7 – Watson OpenScale**
04:15 PM – 04:20 PM	**Lab-8 Presentation**
04:20 PM - 05:00PM	**Lab-8 – Decision Optimization**
05:00 PM - 05:15 PM	**Wrap-up**

IBM Federal and Public Sector Garage
Co-create and co-execute a minimally viable product



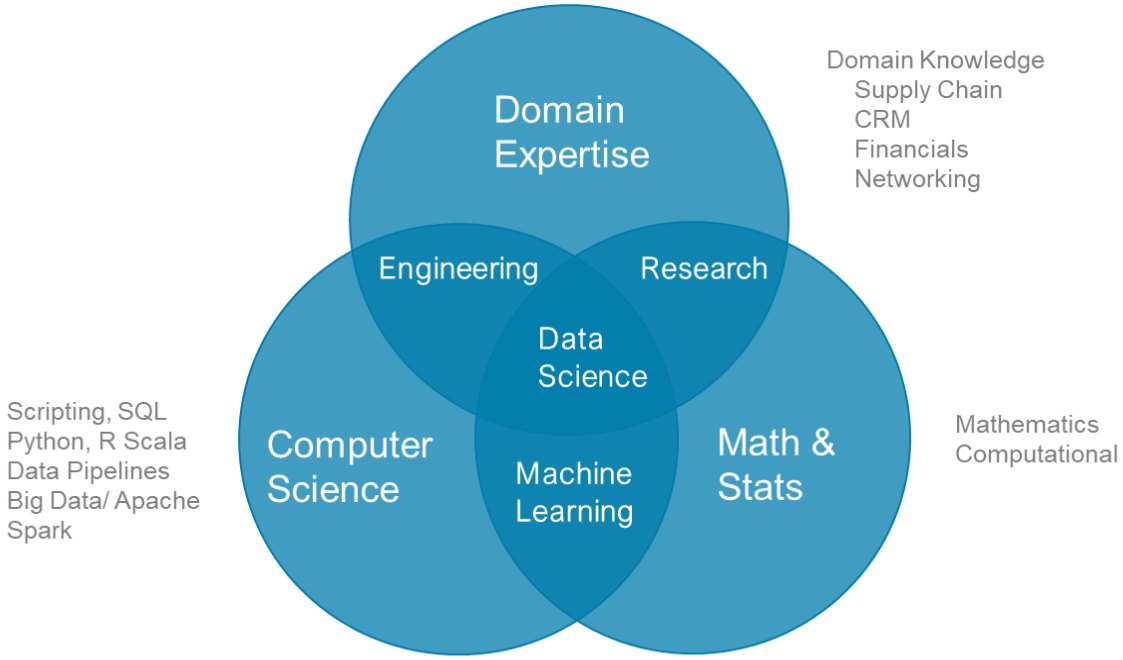
Get started co-creating with the IBM Garage

	Framing	Discovery	Solutioning	Scoping	MVP Build
Sessions	2 - 4 hours	2 hours - 2 days	2 hours - 2 days	2 - 4 hours	3 - 6 weeks
Client outcomes	• **Understand** the strategy • **Determine** business / technology initiative(s) to focus on • **Align** stakeholders on vision and desired outcome • **Confirm** executive sponsor, product owner, and governance model	• **Understand** target end users • **Understand** 'as-is' context of business and/or technology • **Guide** narrowing focus	• **Diverge** to explore potential solutions • **Converge** to select solution to invest in validating • **Identify** platform / initial technical components to be used • **Develop** roadmap	• **Define** hypothesis to be tested / proof-points to be proven • **Define** scope of MVP • **Identify** resources needed to build MVP	• **Build** MVP that leverages IBM hybrid cloud technologies • **Define** a secure minimum viable architecture that mitigates risk • **Set up** cloud platform and automation • **Build** skills and evolve culture through pairing • **Create** an implementation roadmap for a hybrid, multi-cloud platform and DevOps adoption that leverages IBM hybrid cloud
Approach	• Business landscape • Initiative exploration • Vision definition • Opportunity canvas	• Process mining • End-user research • Technical discovery • Data assessment • Modernization assessment	• Visioning • Generating big ideas • 'Just enough' architecture • Rapid prototyping • Identify accelerators • Platform initiation	• Hypotheses definition • MVP definition • Data required • End user validation needed	**WORKLOAD** **PLATFORM**

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab Presentations**

What is Data Science?

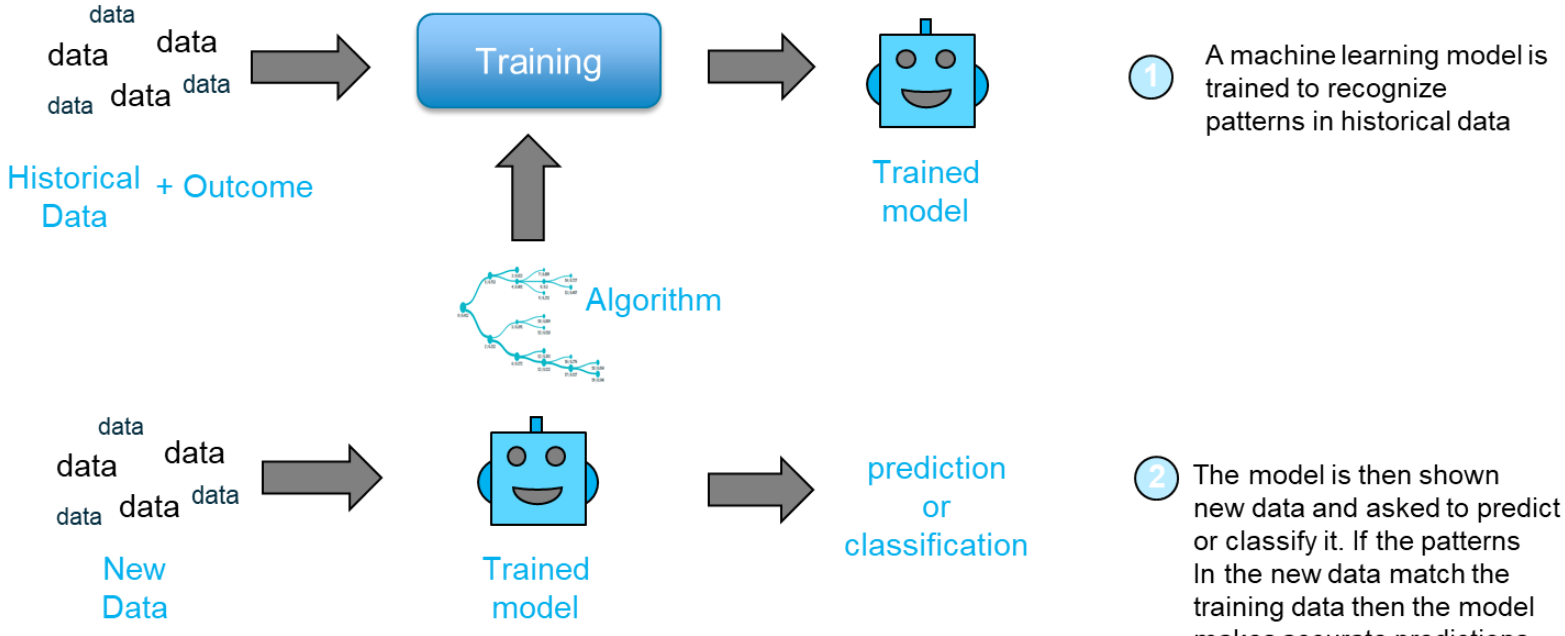


Data Science Projects Require Multiple Skills

Modified from Drew Conway's Venn Diagram

What is Machine Learning?

*"Computers that learn without being **explicitly programmed**"*

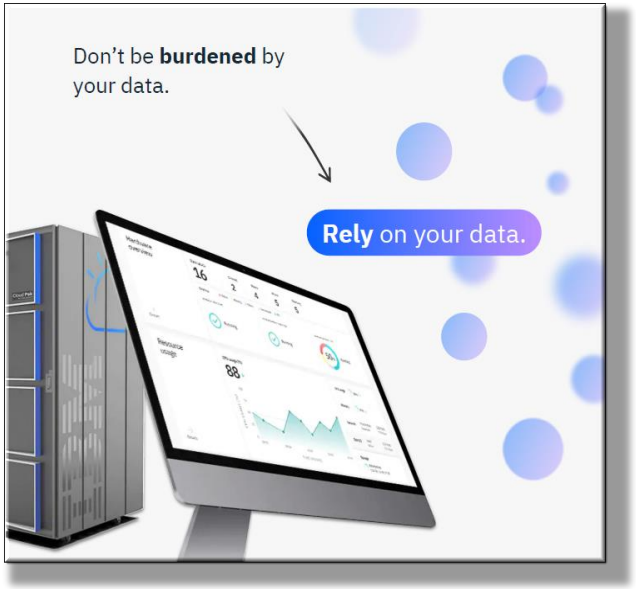


IBM Cloud Pak for Data

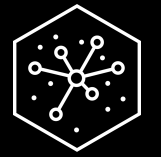
IBM Cloud Pak for Data is a single unified, integrated platform which helps to simplify the collection, organization and analysis of data.

With it, enterprises can turn data into insights through an integrated cloud-native architecture.

IBM Cloud Pak for Data is extensible and easily customized to unique client data and AI landscapes through an integrated catalog of IBM, open source, and third-party microservices.



Considerations for Cloud Pak for Data

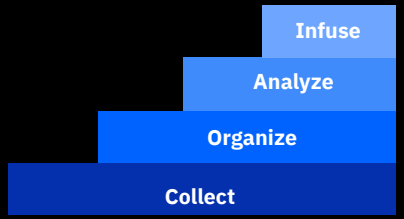


- Integrated Multi-modal platform
  - Use tool of choice and collaborate via project entities
  - Code/Click Options
  - All Analytics – Dashboard, Predictive, Prescriptive
  - All Data
  - Seamless user experience
- Hybrid Cloud
  - Cloud native architecture
  - Cloud agnostic – any vendor cloud or 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

IBM Cloud Pak for Data

Unified, modular, deployable anywhere

The AI Ladder



App Developers and SREs| Business Partners | Data Engineers | Data Stewards | Data Scientists | Business Users

Integrated User Experience

Extensible: APIs, partner ecosystem, and solutions

Collect

- Provision SQL and NoSQL databases
- Create Connections
- Data virtualization
- ...

Organize

- Data quality and classification
- Policies and rules
- Data cataloging
- Self-service discovery & search
- Data transformation
- ...

Analyze and Infuse

- Business reporting
- Data science and visualization
- AI lifecycle automation
- AI Apps
- Industry accelerators
- ...

Core Services

- User access management
- Security contexts and RBAC
- Volume management
- Monitoring and metering
- Service provisioning
- Operators
- Diagnostics
- Backup and migrate

Red Hat OpenShift

IBM Cloud | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

Cloud Pak for Data Deployment Options

- Cloud Pak for Data as a Service
  - Managed offering provided on the IBM Cloud
  - 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

IBM Cloud Satellite

Build Faster. Securely. Anywhere

Consume IBM Cloud services anywhere.

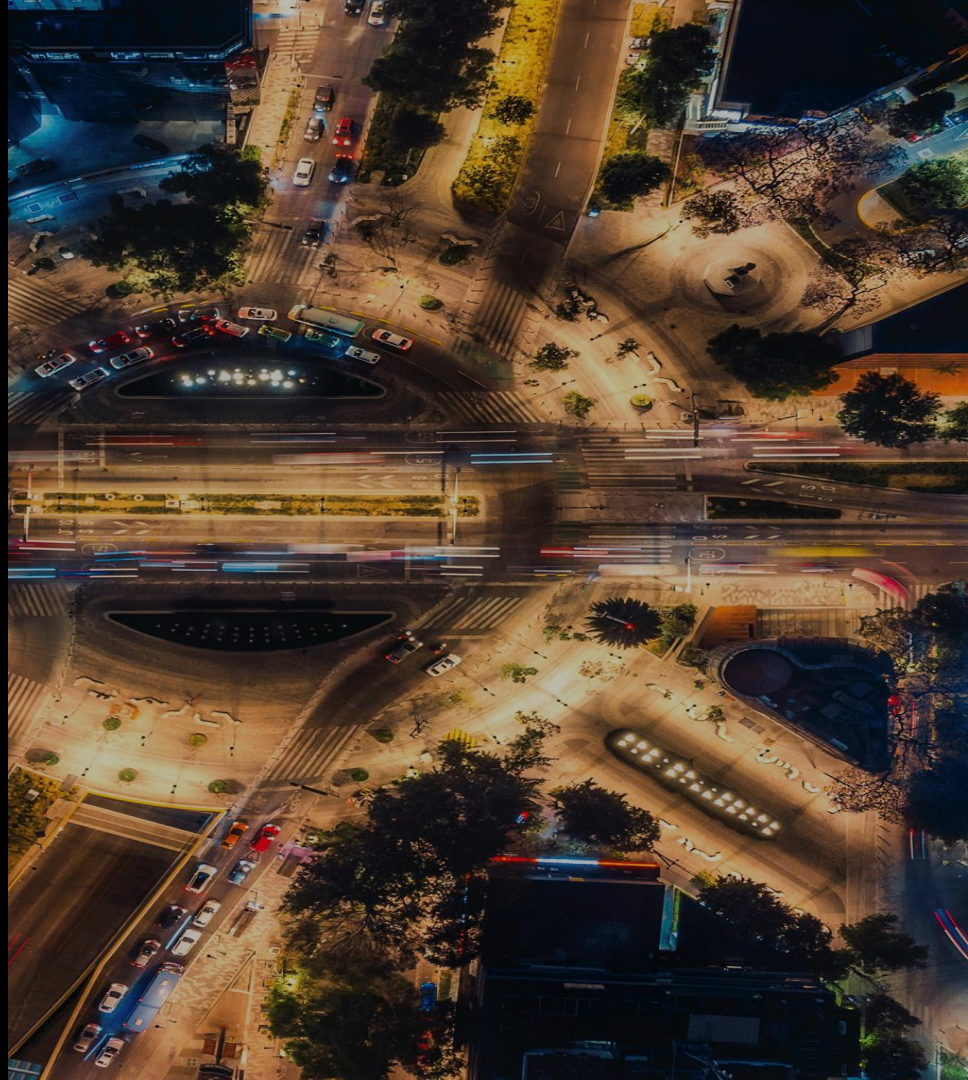
- On prem, at the edge or on other public clouds

Flexible Infrastructure options:

- Meet you where you are with existing infrastructure, integrated appliances or your public cloud accounts

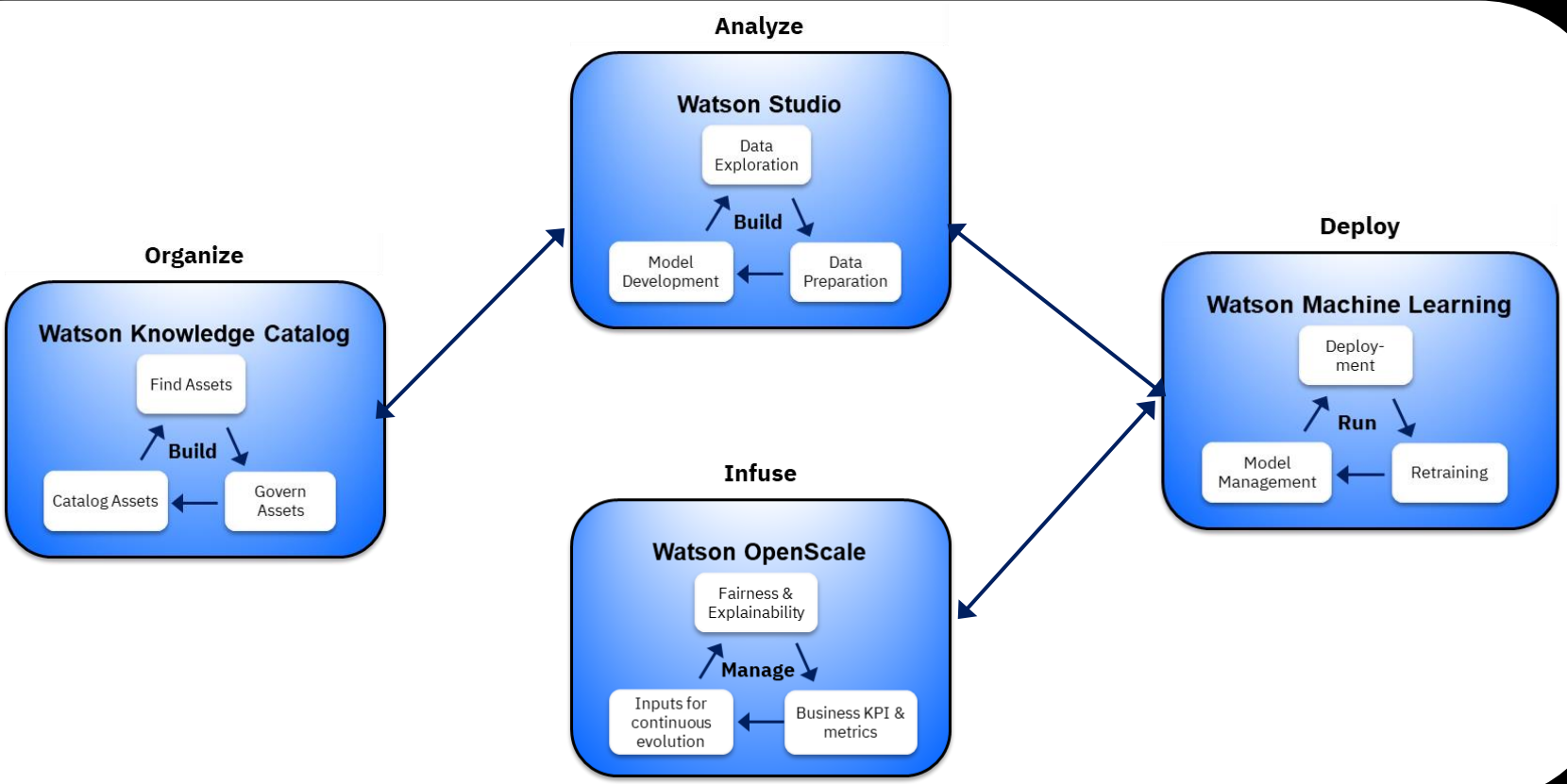
Industry Optimized including FS Cloud :

- Controls strong enough for banks and regulated industries coming 2H2021



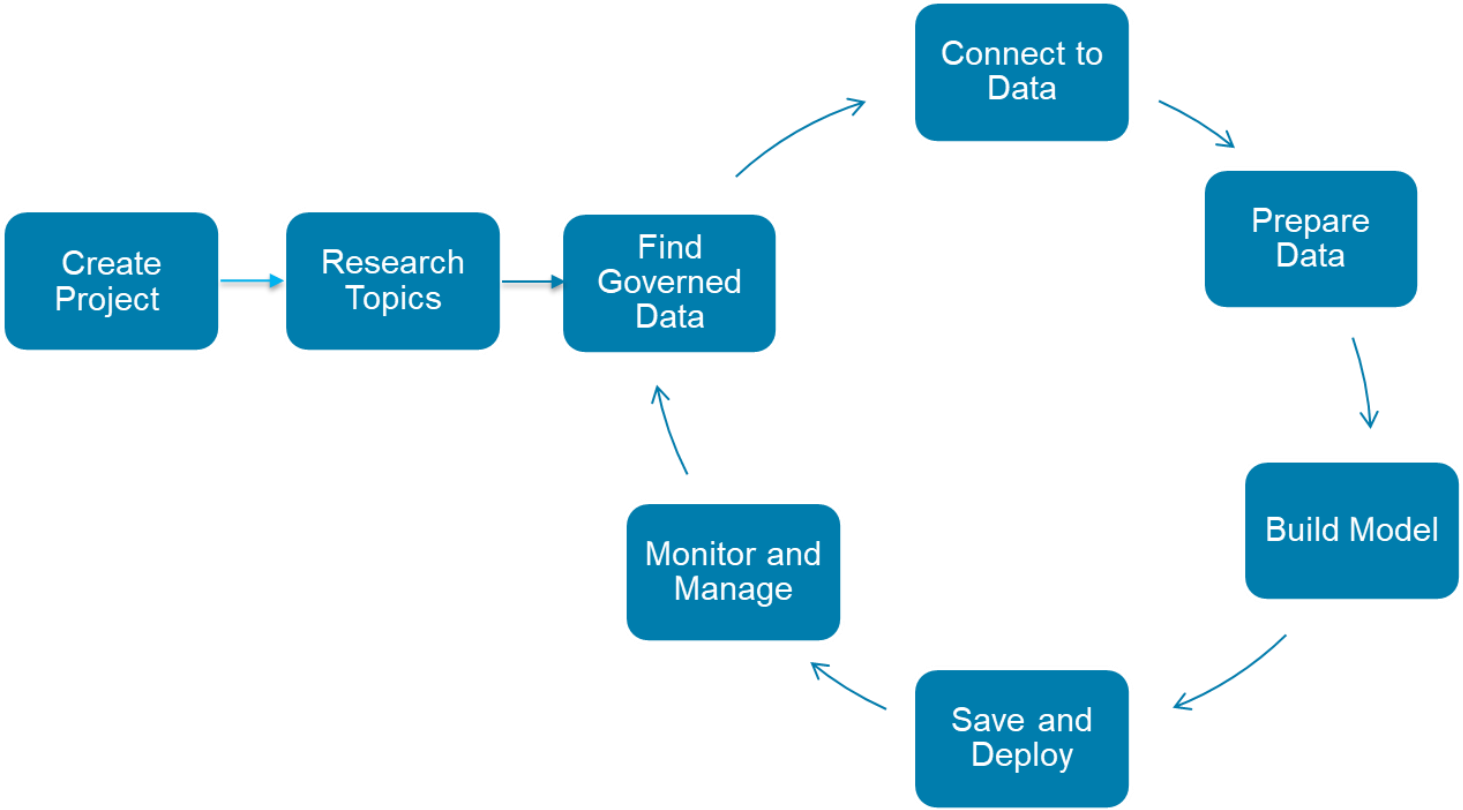
Organize, Analyze, Deploy, Infuse

Cloud Pak for Data services used in today's labs



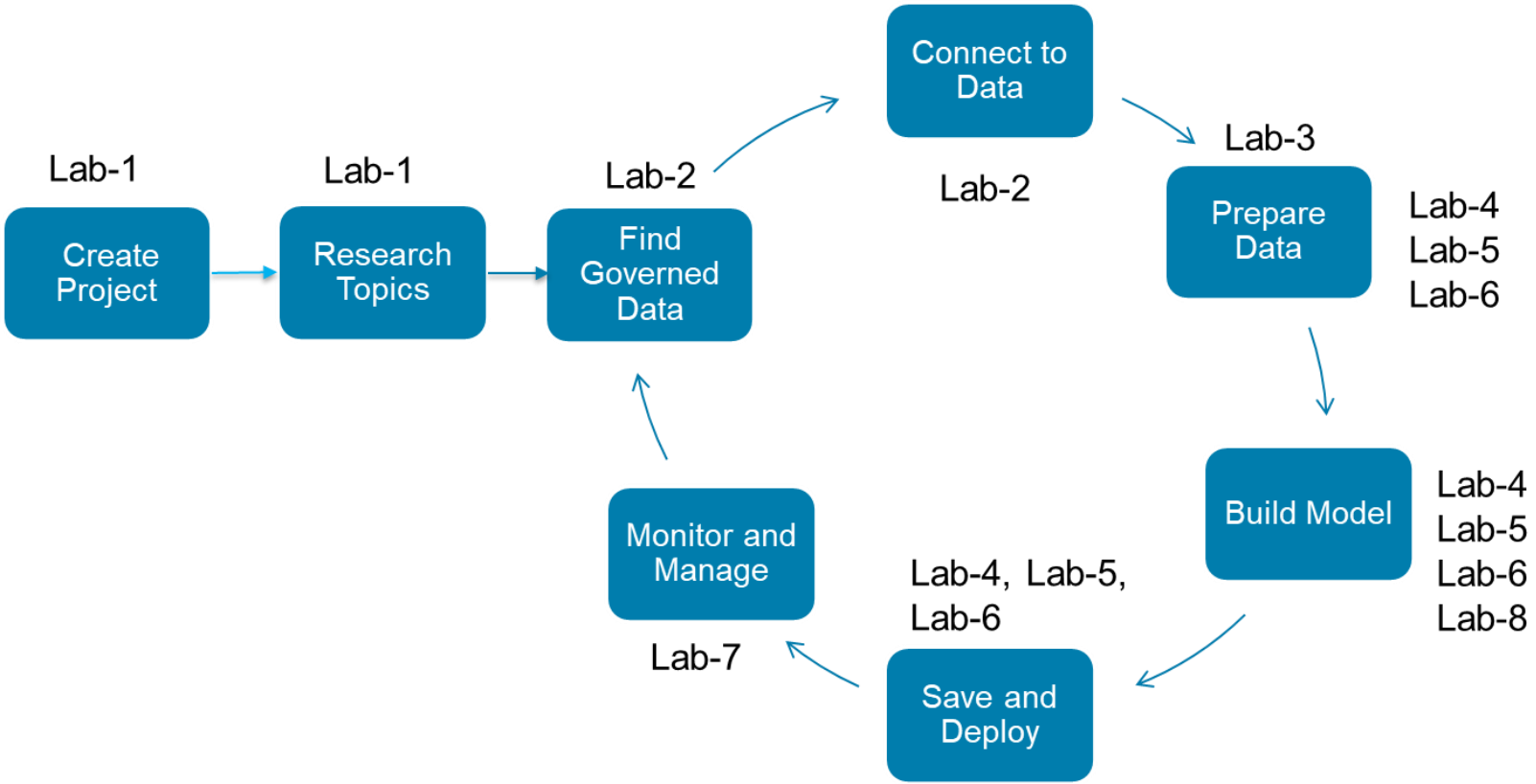
Cloud Pak for Data Supports the Data Science Lifecycle

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



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

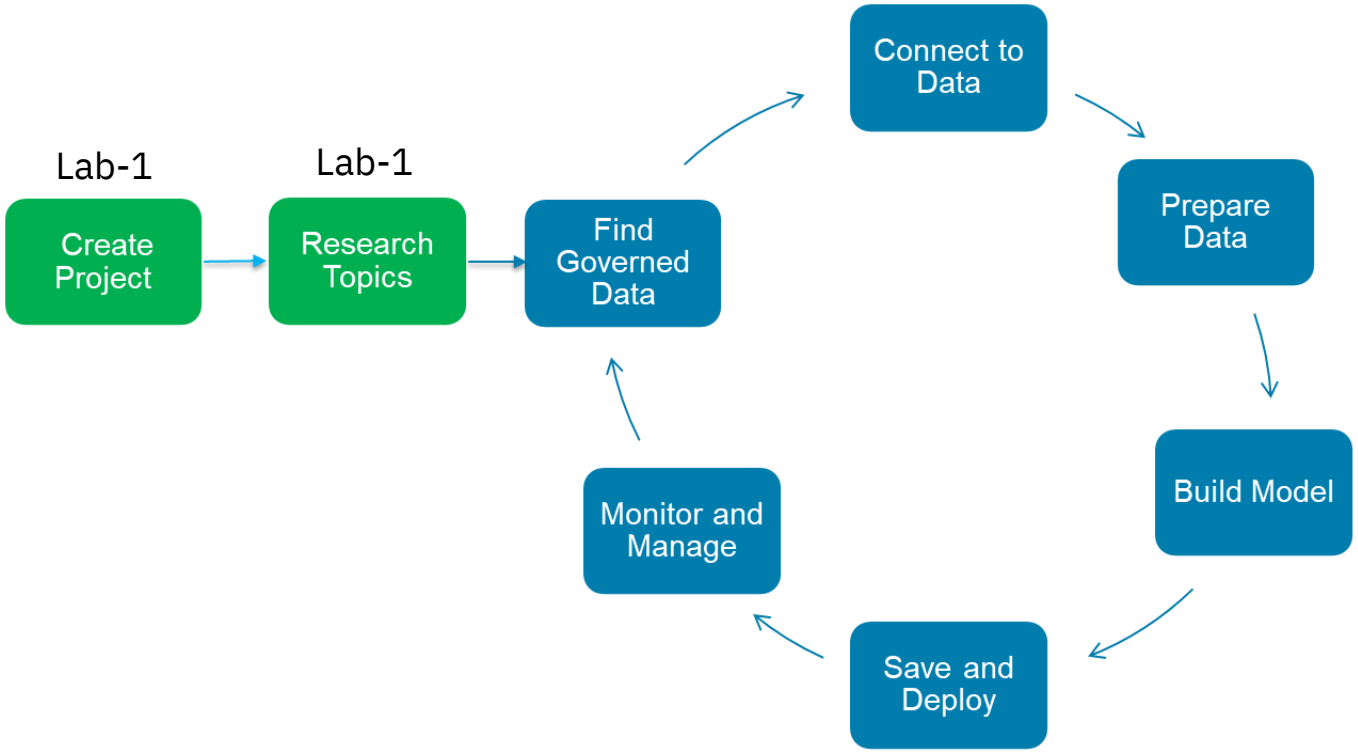


Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 1 Presentation**
  - **Watson Studio Project**
  - **Watson Studio Gallery**

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

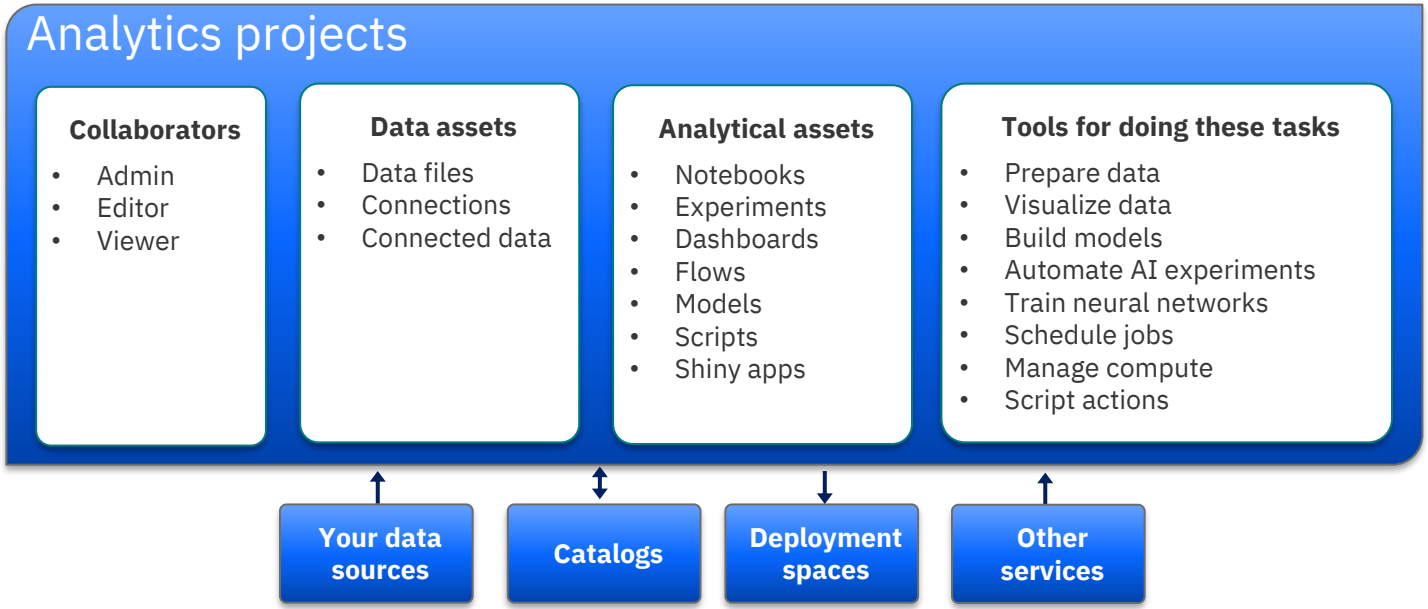


Watson Studio Projects

Making Data Science a Team Sport

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

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



Watson Studio Gallery

Built-in learning to get started

- The Gallery includes sample projects, notebooks, and data sets
- Copy notebooks or Data Sets into projects
- Instantiate sample projects
- Continuously updated

Lab-1: Objectives

Become familiar with Watson Studio projects and gallery by:

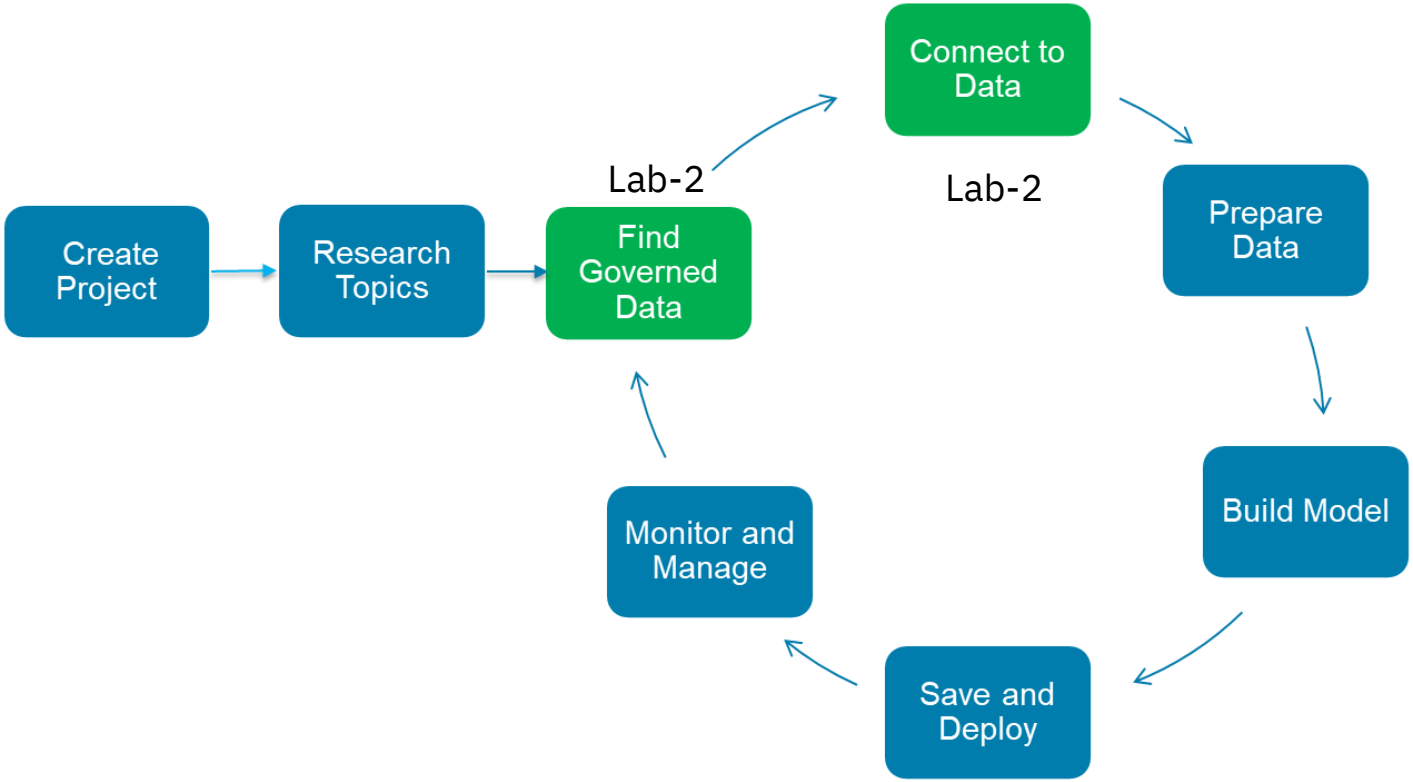
- Creating a Project
- Creating a Watson Machine Learning service instance and associate it with the project
- Adding a collaborator to the project
- Researching topics by searching the Gallery

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 2 Presentation**
  - **Watson Knowledge Catalog**

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

Unlock tribal knowledge and unleash knowledge workers

- **Find** data (structured, unstructured) and AI assets (e.g., ML/DL models, notebooks, dashboards) 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
- Business Glossary

Connection Features

Connect to your data sources

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

Lab-2: Objectives

Become familiar with the Watson Knowledge Catalog by:

- Creating a governed catalog
- Adding a member to the catalog
- Adding Data Assets and Connections to the catalog
- Pushing the Data Assets and Connections to the project set up in Lab-1
- Searching the catalog
- Editing/Reviewing/Profiling a structured Data Asset
- Reviewing access control features

Lab Use Case: Female Human Trafficking

Used in Labs 2-6

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.

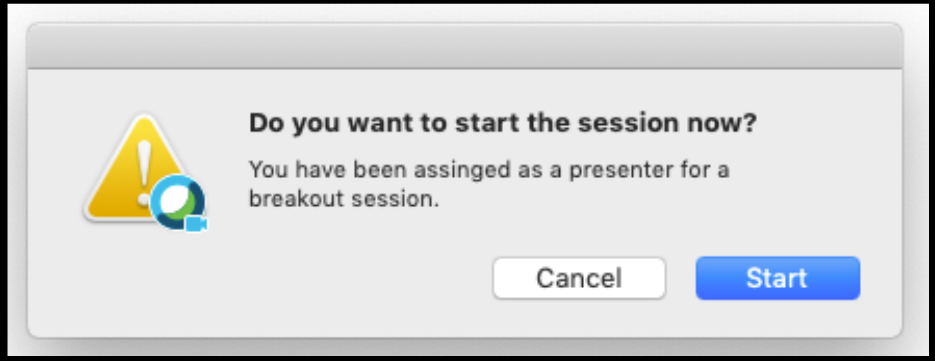
Lab Data

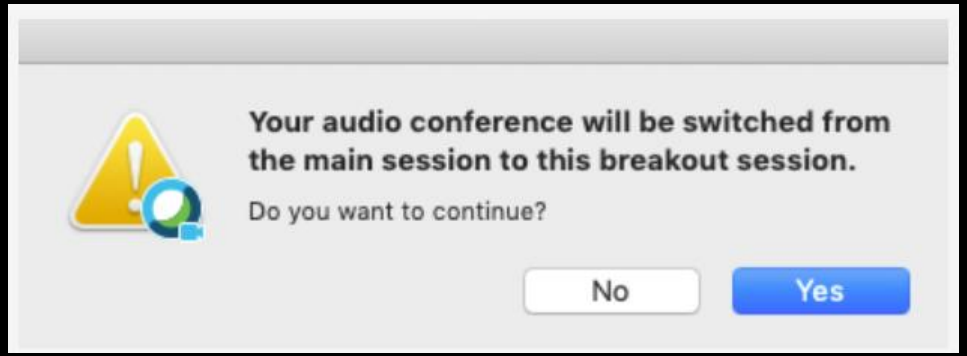
Field	Description
UUID	Hash-based unique identifier
VETTING_LEVEL	Analyst vetting status : 100- PENDING, 10 – HIGH, 20 – MED, 10 - LOW
NAME	Person name
GENDER	Person Gender
AGE (SPSS Modeler)	Person age at time of travel
BIRTH_DATE (Notebook)	Person birth date
BIRTH_COUNTRY	Person full birth country
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
COUNTRIES_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.

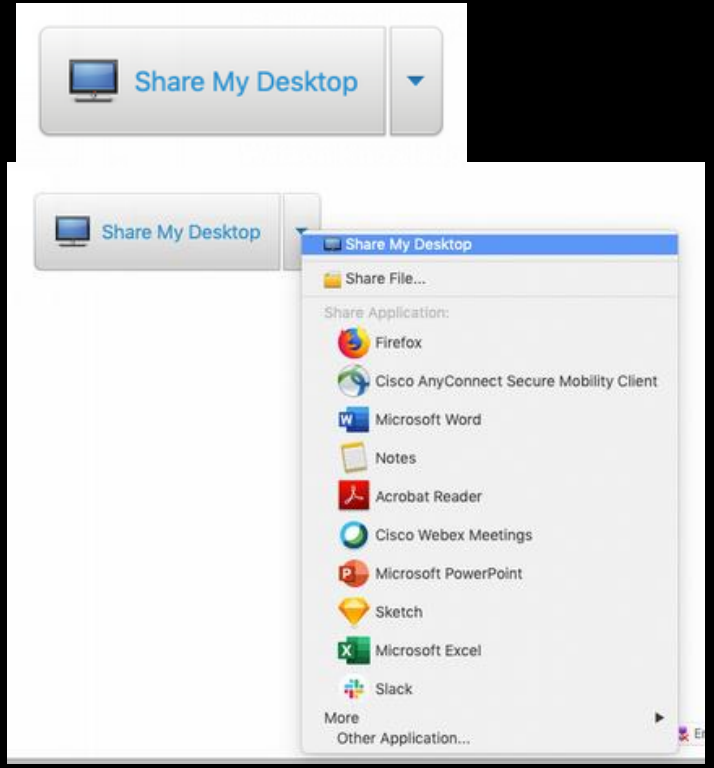
Target

Features

Breakout Rooms





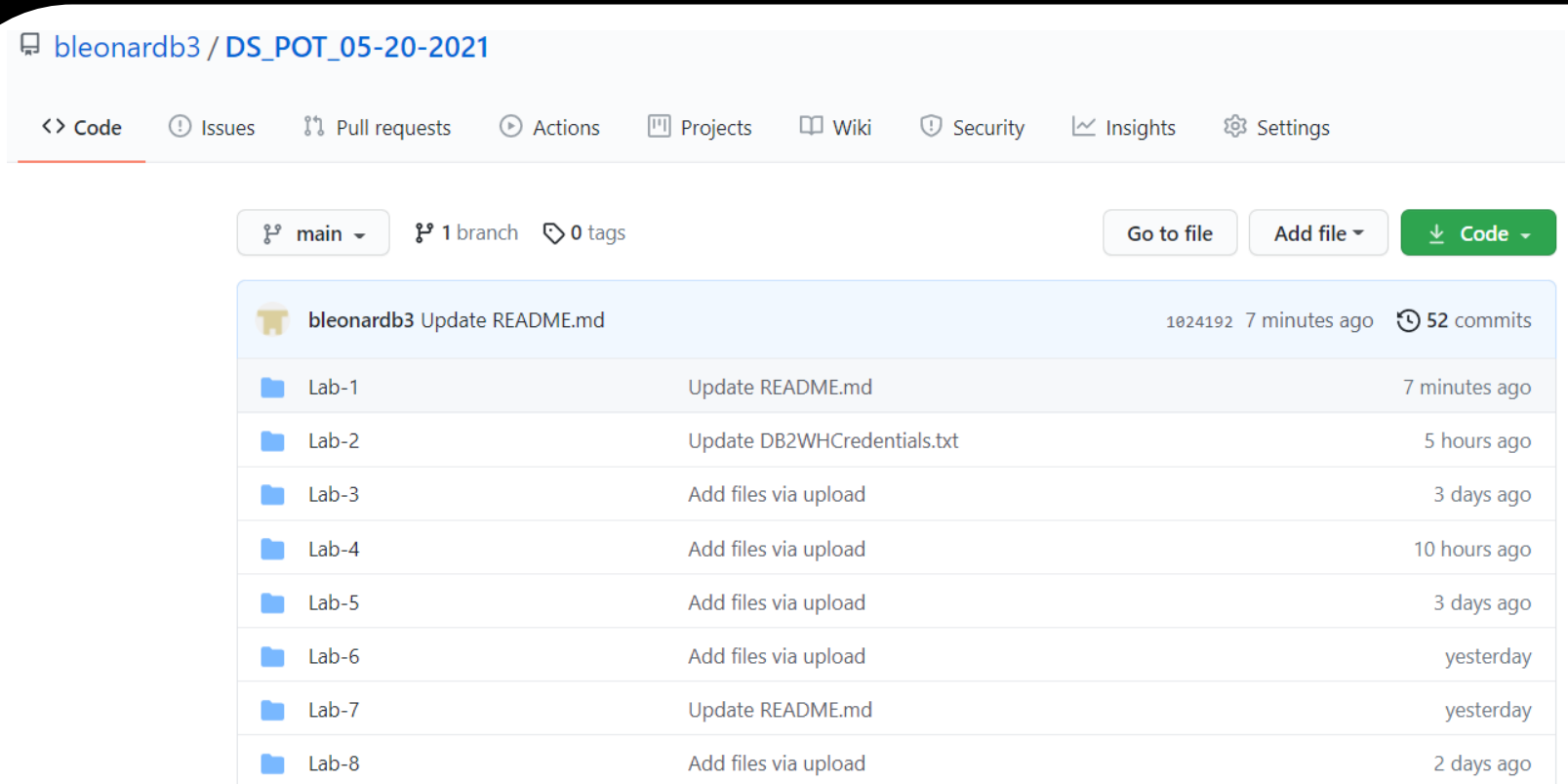


Note: you will need to un-mute when you join the breakout room

Lab Tips

- Labs are in www.github.com/bleonardb3/DS_POT_05-20-2021 repository.
- You will need an IBM Cloud account that has been activated for Watson Studio use. In case, you didn't have a chance to signup, the instructions are in the github repo.
- 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.
- Use Firefox or Chrome browsers. Do not use Internet Explorer or Edge as the browser. For Mac users do not use Safari.

Github Repository



Github Repository

Lab-1: Setup Environment

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

Instructions:

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

Instructions.

Cloud Pak for Data

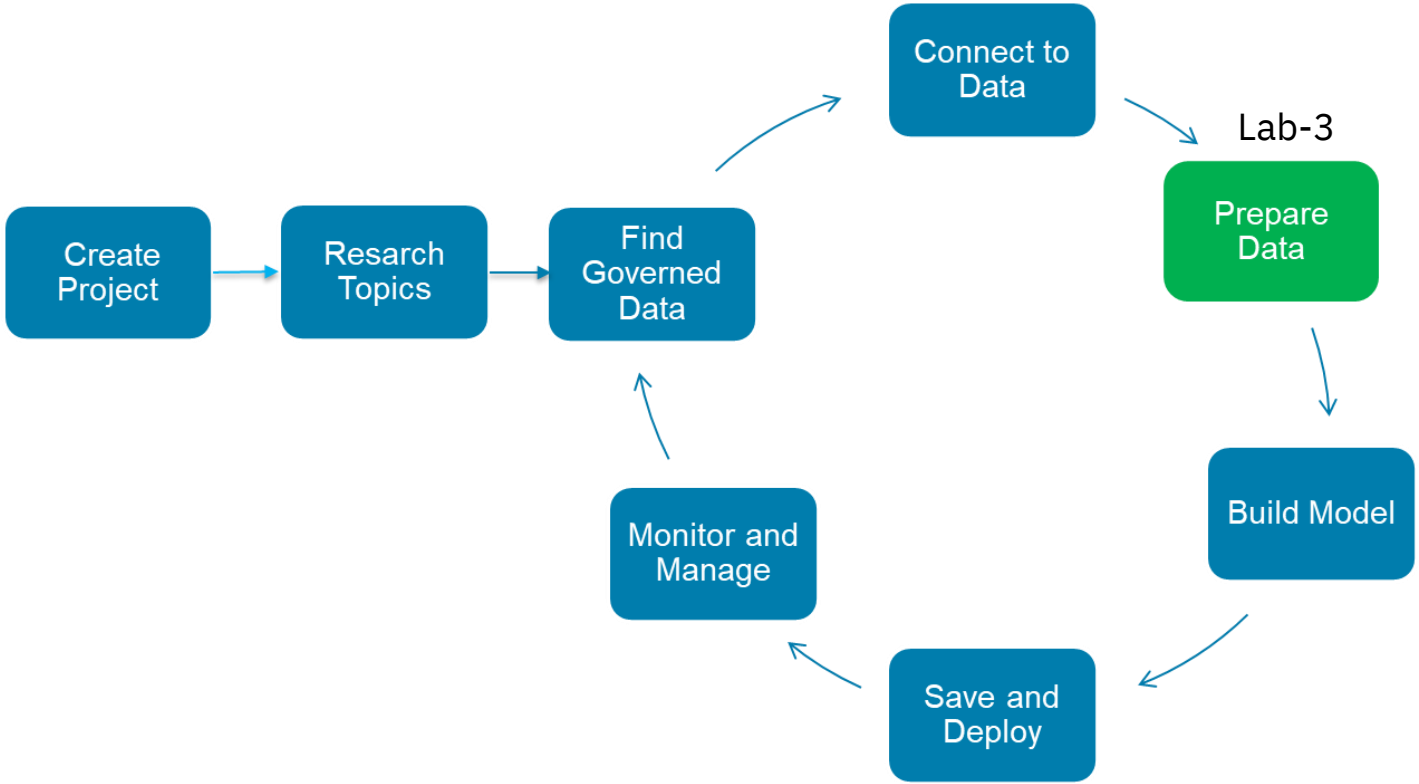
Please work on Lab-1 and Lab-2.
We will return at 11:30 am EST
to introduce the next 2 labs.

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 3 Presentation**
  - **Data Refinery**

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



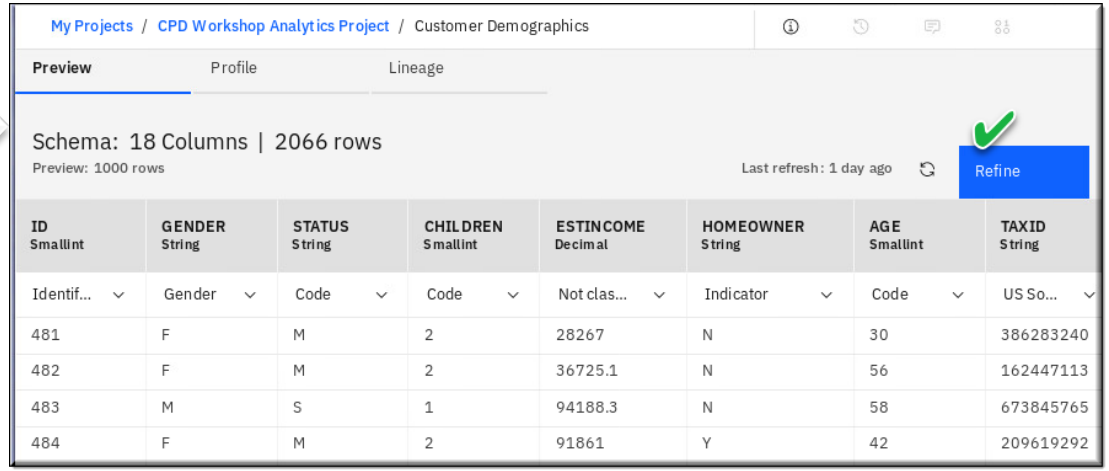
Data Refinery

Making Data fit for use

Refine can cleanse and shape tabular data with a graphical flow editor using functions and logical operators.

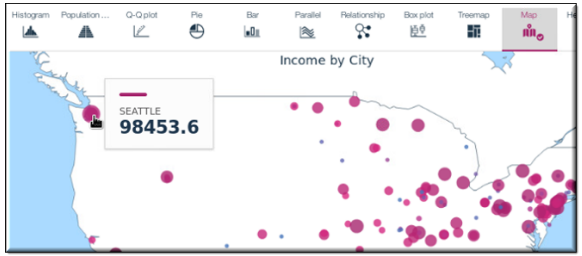
Use it to remove data that is incorrect, incomplete, improperly formatted, etc.

Shape the data by filtering, sorting, combining or removing columns. You can create a Data Refinery flow as a set of ordered operations on the data to run repeatedly any time.



ID Smallint	GENDER String	STATUS String	CHILDREN Smallint	ESTINCOME Decimal	HOMEOWNER String	AGE Smallint	TAXID String
Identif...	Gender	Code	Code	Not clas...	Indicator	Code	US So...
481	F	M	2	28267	N	30	386283240
482	F	M	2	36725.1	N	56	162447113
483	M	S	1	94188.3	N	58	673845765
484	F	M	2	91861	Y	42	209619292

Data Refinery also includes a graphical interface to profile data to validate it with 20+ customizable charts that give perspective and insights into the data.



Lab-3: Objectives

Become familiar with the Data Refinery by:

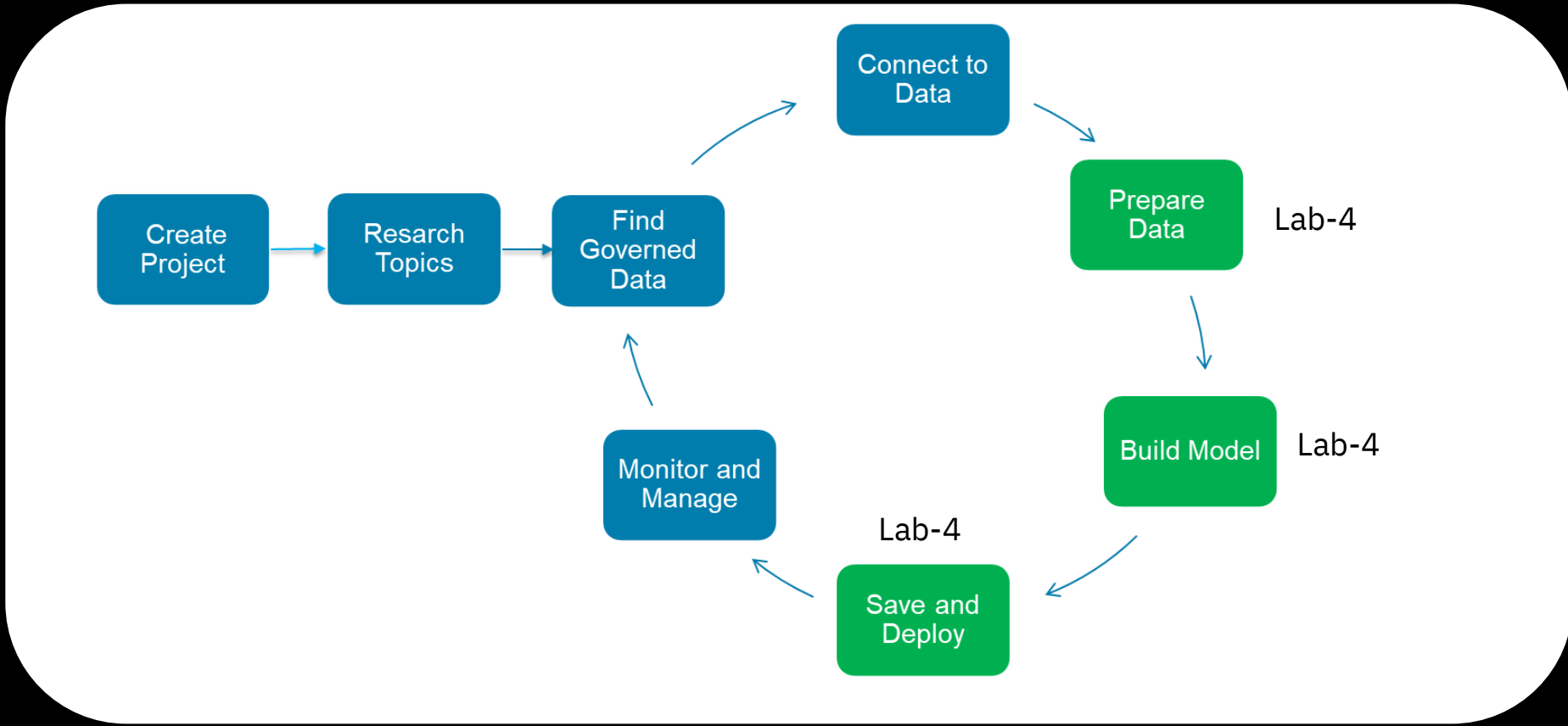
- Creating a new Data Flow
- Profiling the data
- Visualizing the data to gain a better understanding
- Preparing the data for modeling
- Running the sequence of data preparation operations on the entire data set.

Presentation Outline

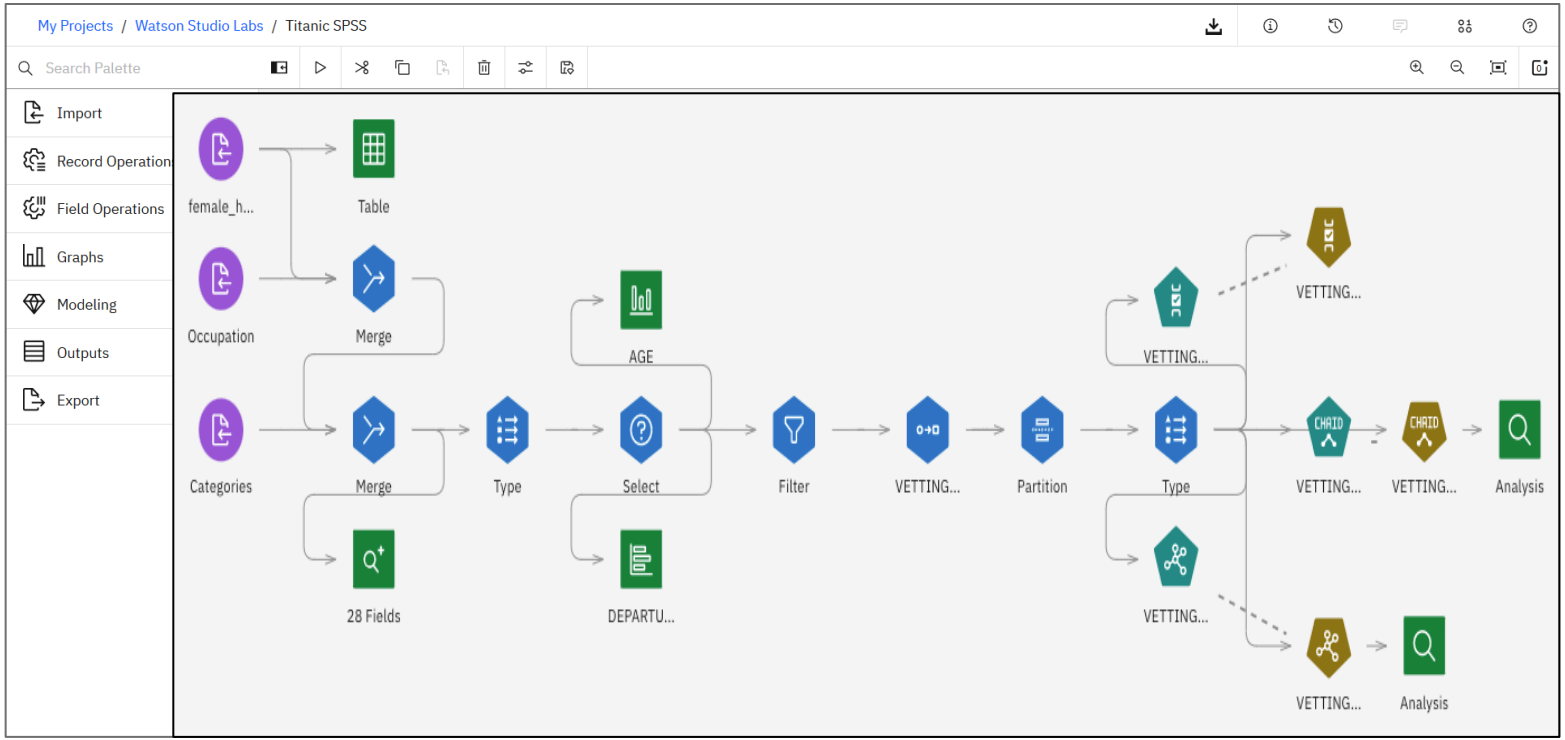
- **IBM Cloud Pak for Data Overview**
- **Lab 4 Presentation**
  - **SPSS Modeler**

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



SPSS Modeler



Lab-4: Objectives

Become familiar with the Watson Studio SPSS Modeler capability by:

- Loading the trafficking data into SPSS Modeler
- Joining the datasets
- Profiling the trafficking data
- Preparing the trafficking data
- Training/Evaluating a machine learning model.
- Saving the model.

Cloud Pak for Data

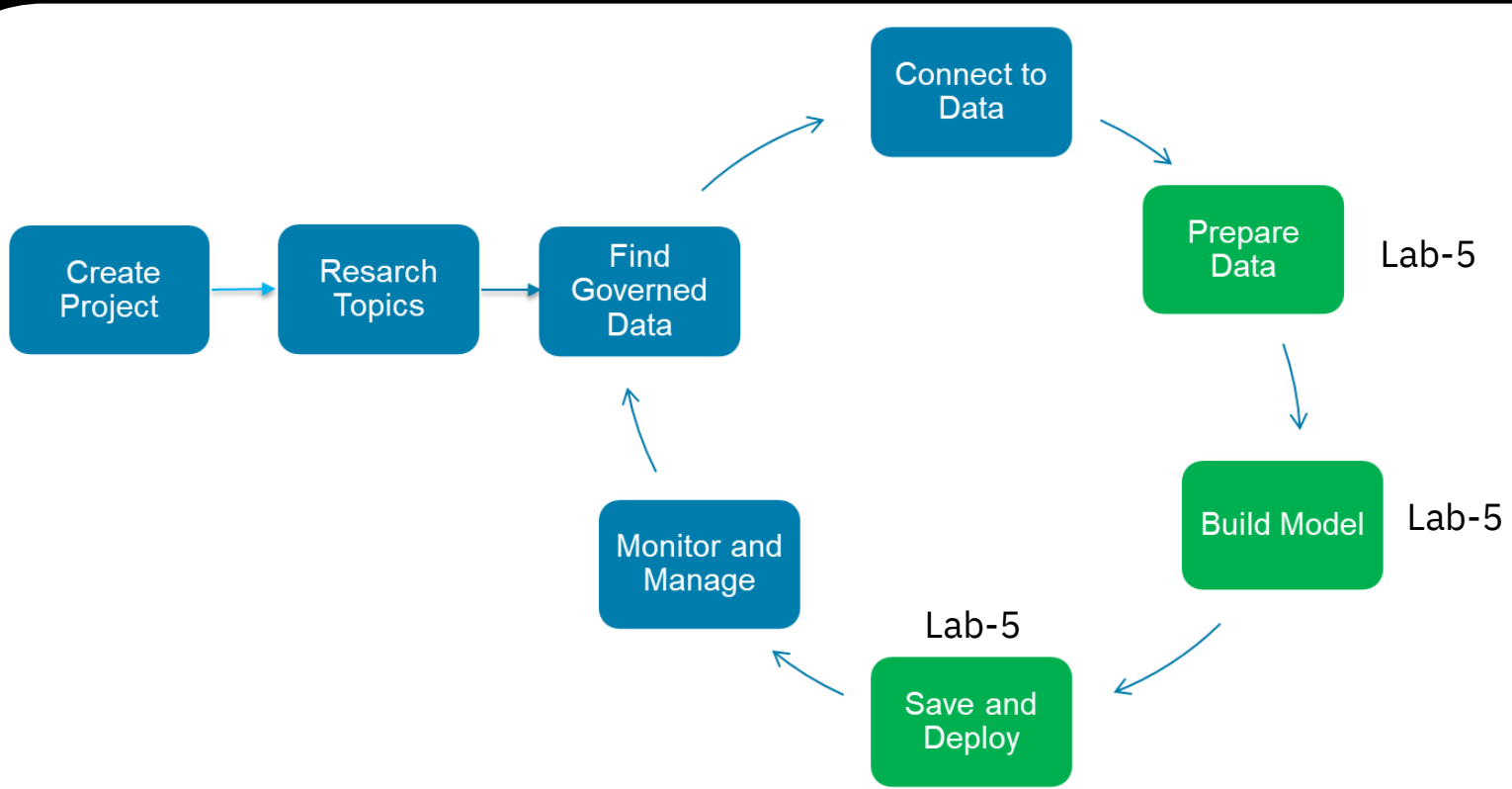
Please work on Lab-3 and Lab-4.
We will return at 1:30 pm EST to introduce the next 2 labs.

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 5 Presentation**
  - **Jupyter Notebook**
  - **SparkML**

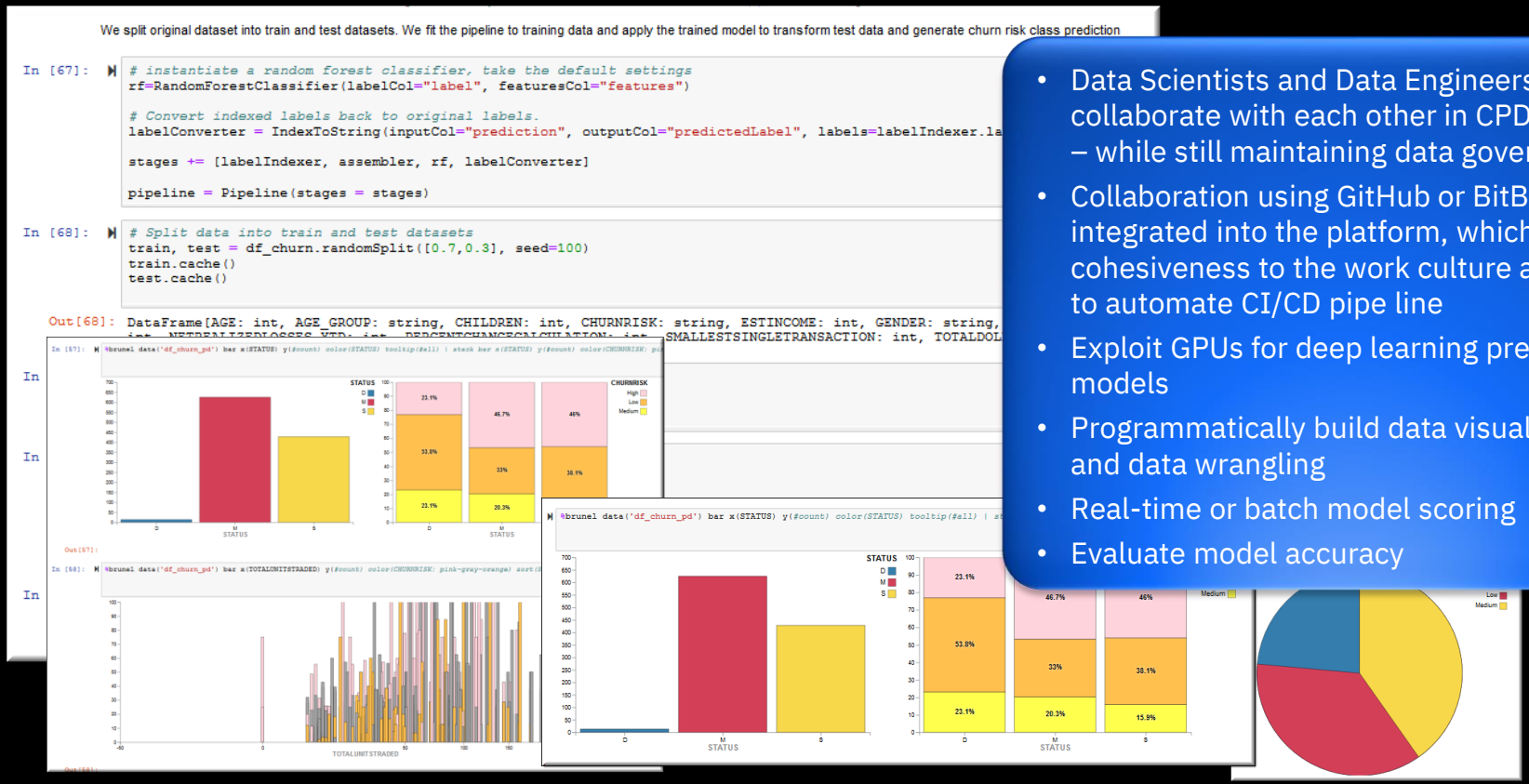
Cloud Pak for Data Supports the Data Science Lifecycle

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



Watson Studio Notebooks

Build Data Science & Machine Learning models



- Data Scientists and Data Engineers collaborate with each other in CPD platform – while still maintaining data governance
- Collaboration using GitHub or BitBucket is integrated into the platform, which brings a cohesiveness to the work culture and helps to automate CI/CD pipe line
- Exploit GPUs for deep learning predictive ML models
- Programmatically build data visualizations and data wrangling
- Real-time or batch model scoring
- Evaluate model accuracy

Spark Overview

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

Lab 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

Lab Flow

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

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: Objectives

Become familiar with the Notebooks and Spark by:

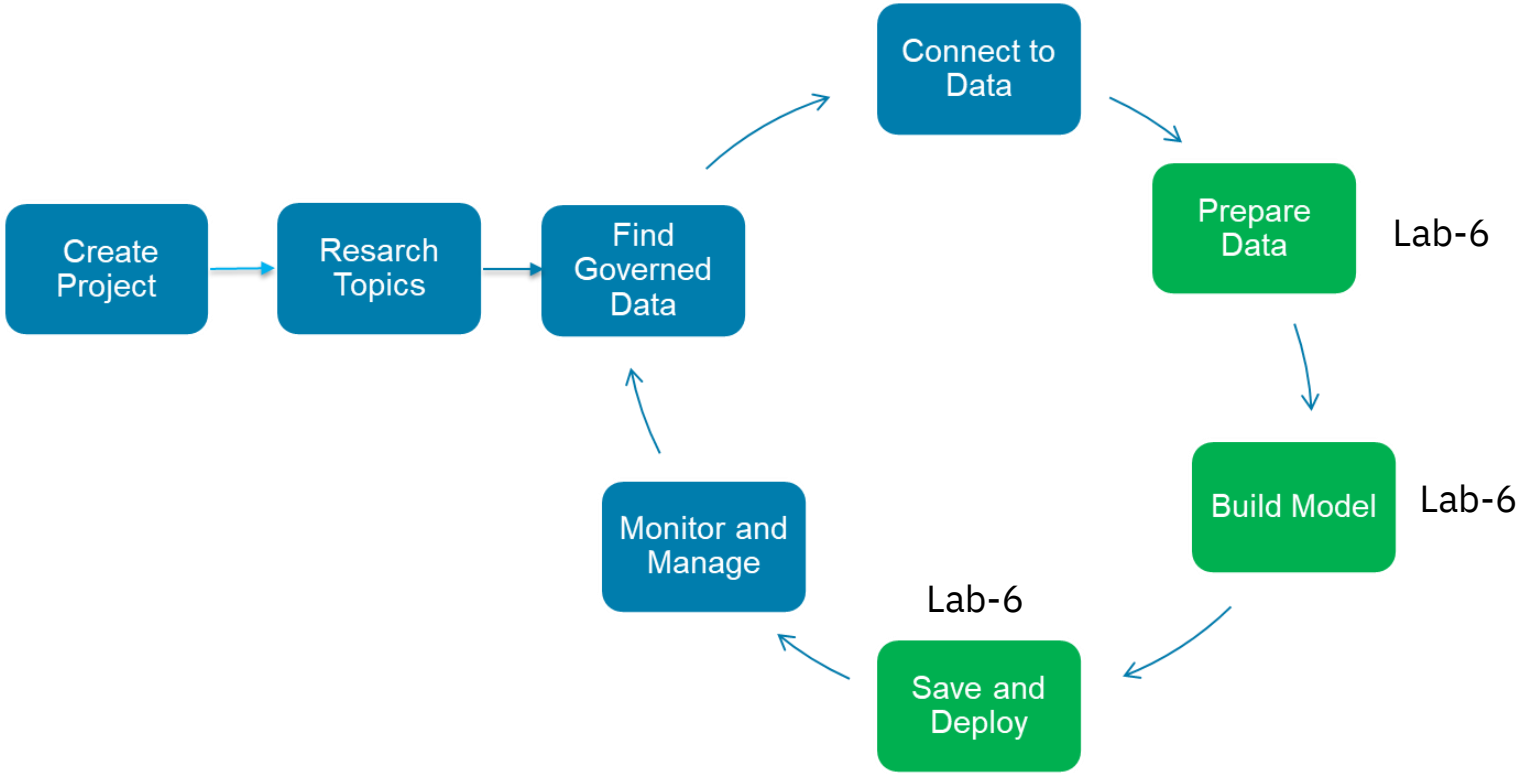
- Creating a project token
- Joining data from three sources.
- Identifying labels and defining transforms
- Declaring a machine learning model.
- Setting up the Pipeline for data transforms and training.
- Training the model.
- Evaluating and showing model results
- Automatically tuning model
- Saving the model to the model repository.

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 6 Presentation**
  - **AutoAI**

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



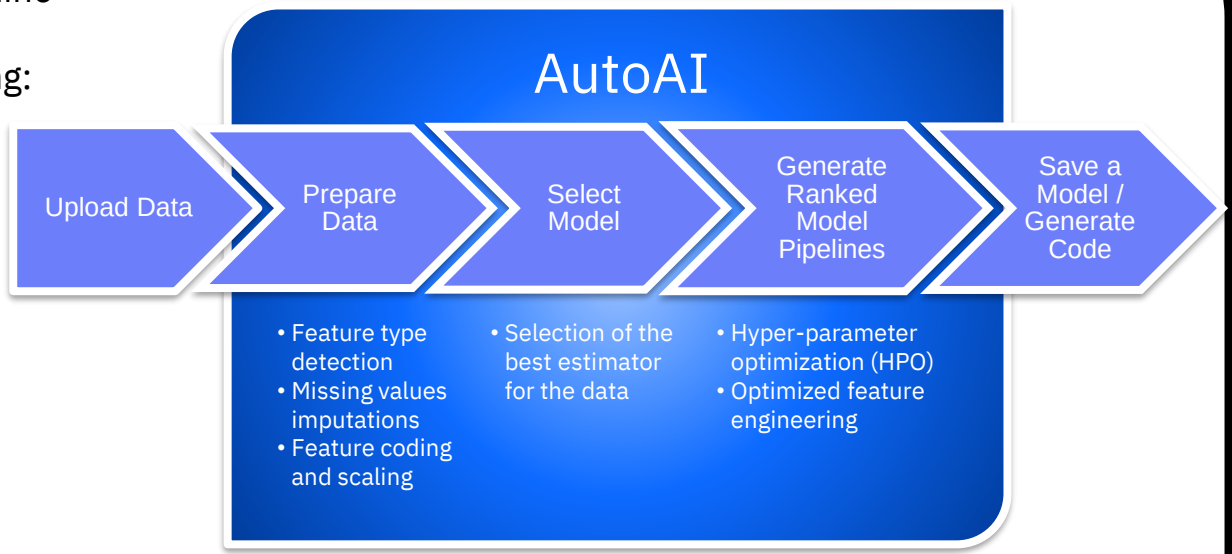
Auto AI
Overview

AutoAI is an award-winning technology that simplifies the Machine Learning model creation and AI lifecycle by automating the following:

- **Data preparation**
- **Model development**
- **Feature engineering**
- **Hyper-parameter optimization**

AutoAI delivers training feedback visualizations for real-time model performance results with:

- **Binary, Multiclass, and Regression support**
- **One-click model deployment**



* AutoAI is enabled with the Watson Machine Learning service install, but it is driven through a Watson Studio Analytics Project

AutoAI
Pipelines

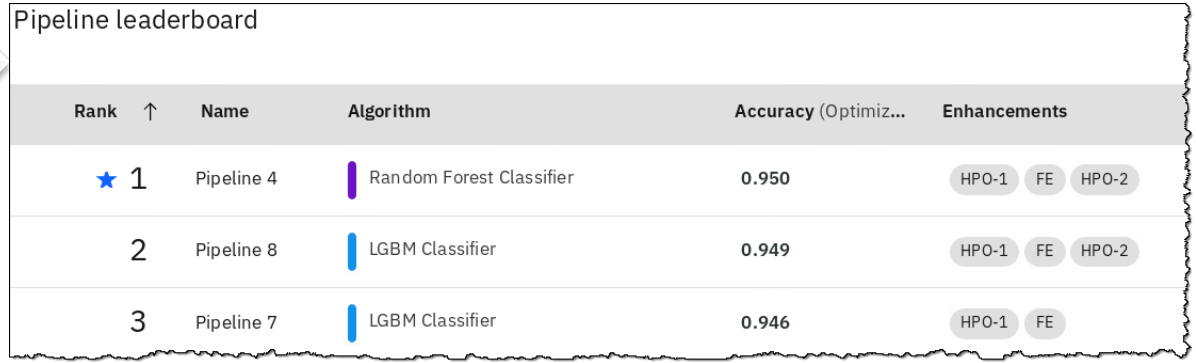
AutoAI pipeline leaderboard

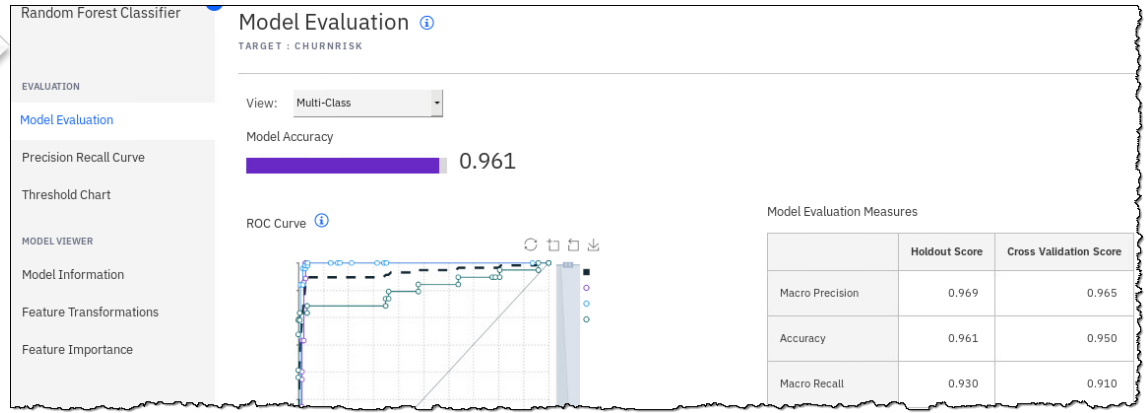
Shows the ranking of the pipelines for each potential model, the higher the better.

After AutoAI completes its model creation steps, you can drill into the pipeline(s) to understand how it came to its conclusion.

Save the pipeline in your project as a:

- **model**
- **notebook**





Lab-6: Objectives

Become familiar with the AutoAI capability by:

- Automatically building a machine learning model
- Saving and Deploying the model
- Testing the model

Cloud Pak for Data

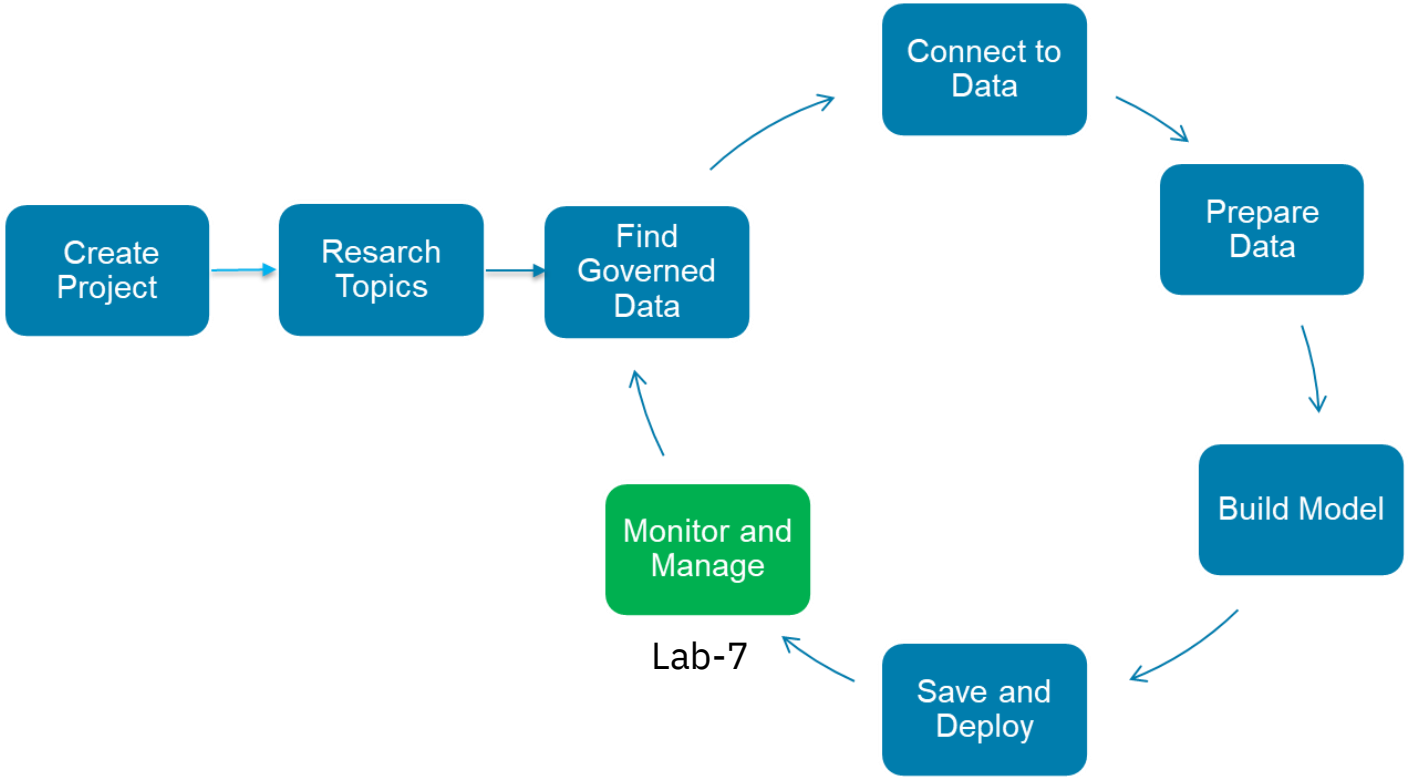
Please work on Lab-5 and Lab-6.
We will return at 4:10 pm EST to introduce the last lab.

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 7 Presentation**
  - **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



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

Monitor AI at Scale

Watson OpenScale

- Operations Dashboard
- 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

Lab-7: Objectives

Become familiar with the Watson OpenScale capability by:

- Configure Watson OpenScale metrics
- Submit Feedback and View Quality Metrics
- Score Data and View Fairness Metrics
- Explain a Transaction.

Cloud Pak for Data

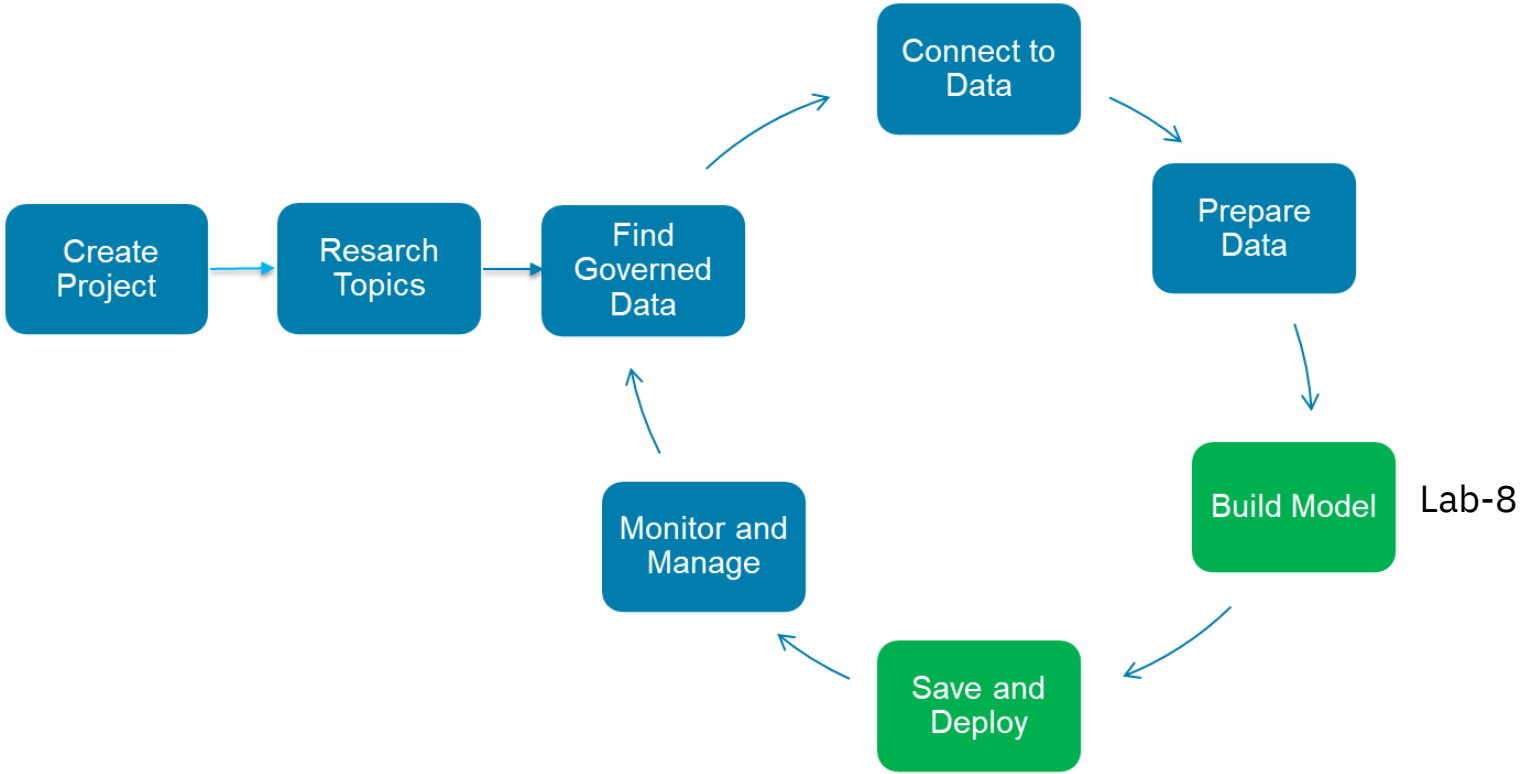
Please work on Lab-7. We will return at 4:15 pm EST to introduce the last lab.

Presentation Outline

- **IBM Cloud Pak for Data Overview**
- **Lab 8 Presentation**
  - **Decision Optimization**

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

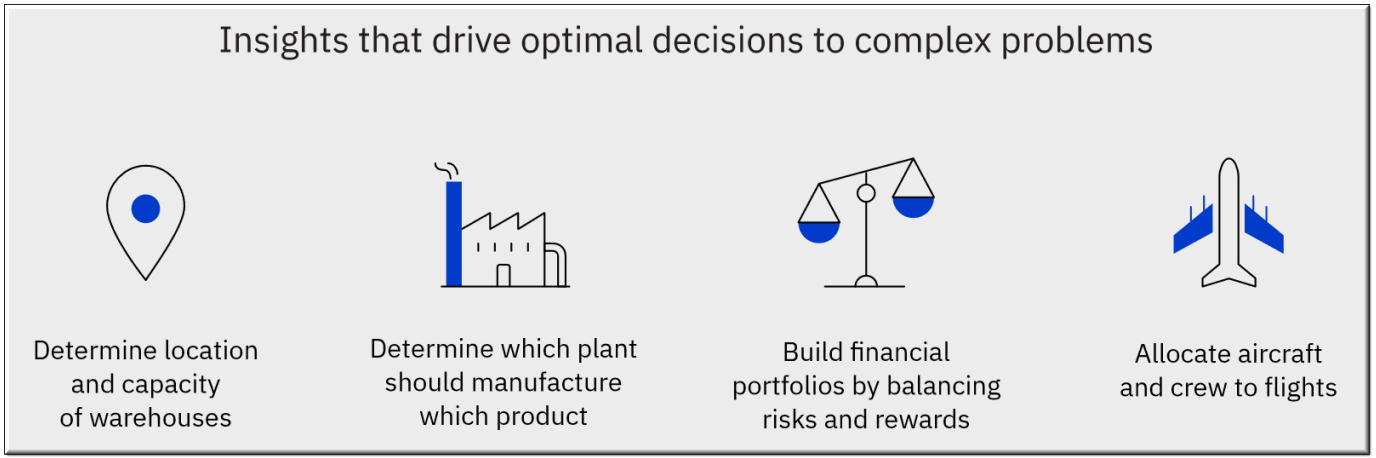


Analyze
Premium Service: Decision Optimization

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.



Lab-8: Objectives

Become familiar with the Decision Optimization capability by:

- Creating a Decision Optimization experiment
- Importing Data into the experiment
- Formulating and running 3 Optimization Scenarios

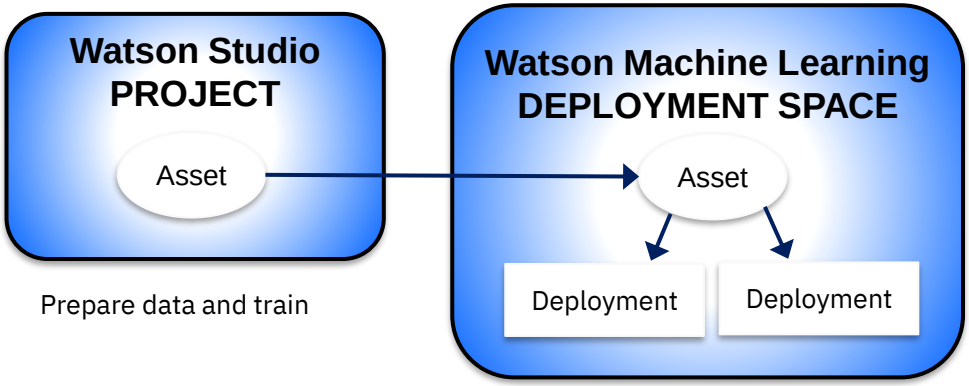
Cloud Pak for Data

Please work on Lab-8. We will return at 5:00 pm EST for a wrap-up.

Deploy
Watson Machine Learning: Deployment Spaces

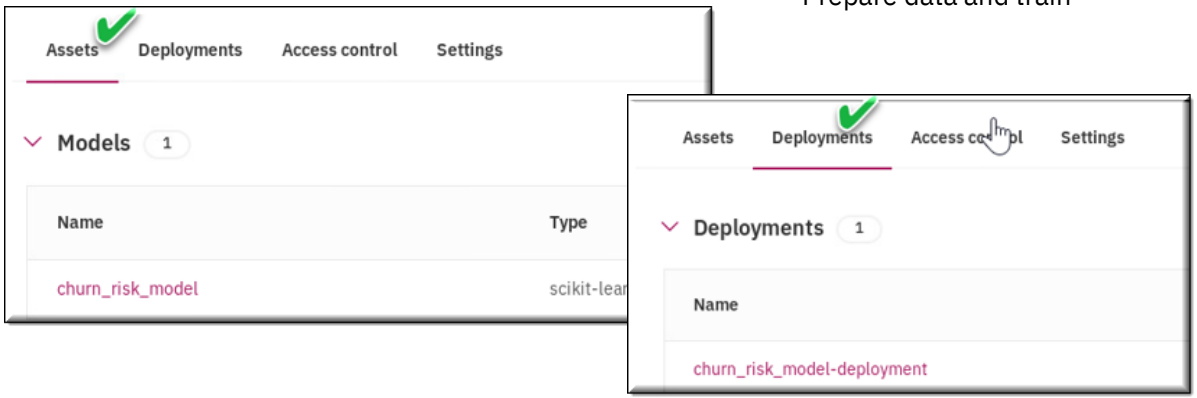
A **Deployment Space** is where you can:

- **Promote and save models**
- **Create the deployments from the models**
- **Find the information you need to score the model and get a prediction back**
- **Embed the deployment in an app so you can interact with it programmatically**



Prepare data and train

Configure, test, deploy and monitor



Deploy
Watson Machine Learning: Deployments

A **Deployment** is the last stage of the model development work. It means you put the model into production so that you can pass data to the model and return a score (or prediction).

After deploying a model, you can access the model *endpoint*, which you will need to make the model available for wider use in applications.

There are three type of WML deployments:

- **Online –** Provides an API endpoint needed to access the deployment programmatically to use in an application. Code snippets are provided in a variety of programming languages that illustrate how to access the deployment.
- **Batch –** Processes input data from a file and writes the output to a file.
- **CoreML**

Additional Information

IBM A3 Center
https://www.ibm.com/industries/federal/analytics

Bernard Beekman
beekmanb@us.ibm.com

James Parry
jparry@us.ibm.com

Michael Cronk
michael.cronk@ibm.com

Frank Stein
fstein@us.ibm.com

Thank you for your time today!

Begin your journey to cloud and AI using IBM's Cloud Pak for Data!!

Copyright and trademarks

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
July 2020