End-to-End Data Science using Watson Studio

April 30, 2020

The session starts at 9:00am.

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

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

Bernie Beekman
Michael Cronk
James Parry

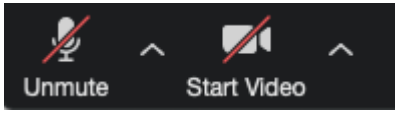
© 2020 IBM Corporation

Agenda

Time	Description
9:00 AM – 10:00 AM	**Overview of Watson Studio** **Lab Orientation 1,2**
10:00 AM – 10:15 AM	**Break**
10:15 AM – 11:45 PM	**Lab 1-2 – Set up Environment, Watson Knowledge Catalog**
11:45 PM 12:45 PM	**Lab Review 1,2 /Lab Orientation 3,4** **Lunch**
12:45 PM – 02:00 PM	**Lab 3,4 – Data Refinery, SPSS Modeler**
02:00 PM – 02:30 PM	**Lab Review 3,4 / Lab Orientation 5,6**
02:30 PM – 03:30 PM	**Lab 5,6 – Machine Learning with SparkML, AutoAI+DevOps**
03:30 PM – 04: 00 PM	**Lab Review 5,6 / Lab Orientation 7**
04:00 PM – 05:00 PM	**Lab 7 – Watson OpenScale**

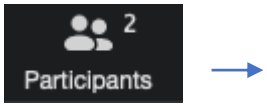
© 2020 IBM Corporation

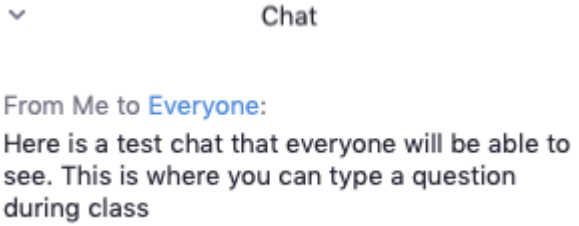
The Zoom Tool Bar

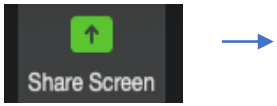


We will stay on mute during the main sessions. You may unmute yourself during breakouts.

We will not use video to preserve bandwidth during our class today.







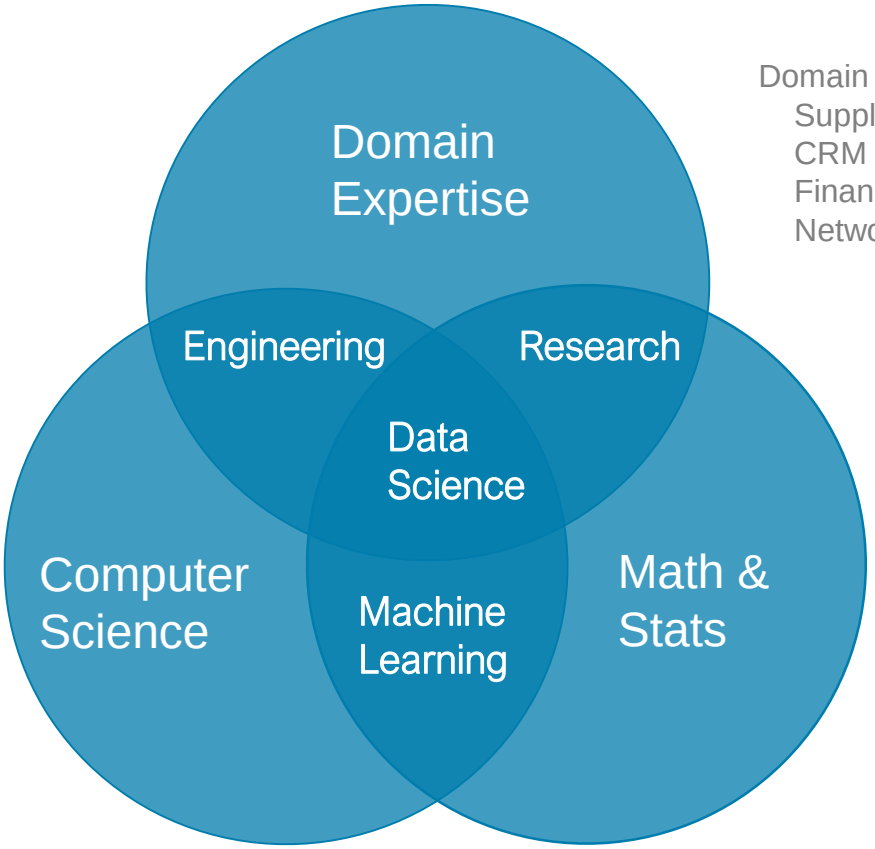
The Share Screen function has been deactivated during the lectures

Outline

- **Data Science Overview**
- **Watson Studio Overview**
- **Lab Overview**

© 2020 IBM Corporation

What is Data Science?



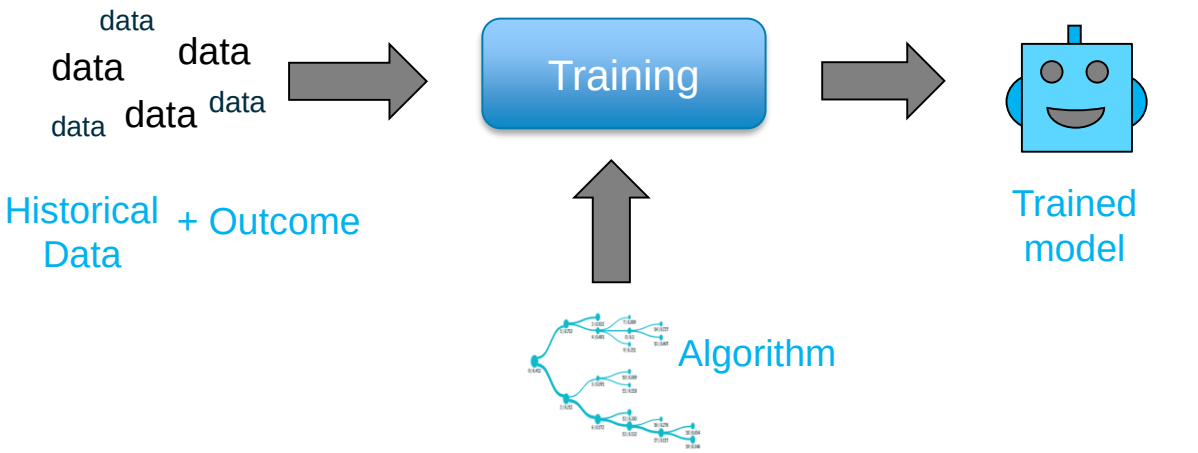
Data Science Projects Require Multiple Skills

Modified from Drew Conway's Venn Diagram

© 2020 IBM Corporation

What is Machine Learning?

"Computers that learn without being explicitly programmed"



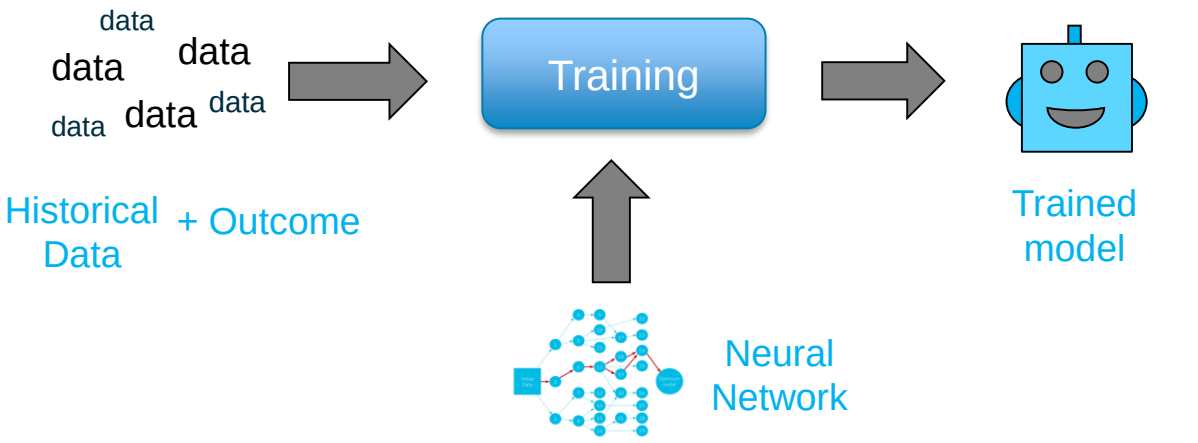
1. A machine learning model is trained to recognize patterns in historical data

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.

© 2020 IBM Corporation

What is Deep Learning?

"Computers that learn without being explicitly programmed"



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.

© 2020 IBM Corporation

IBM takes an Enterprise Approach to Data Science

- Freedom of Choice
  - Choose programming languages, open source libraries, IBM value-add capabilities
  - Code/Click
  - Machine Learning/Deep Learning/Decision Optimization.
  - All Data
- Operationalize Machine Learning
  - Manage complete ML lifecycle – Build, Deploy, Manage, Scale, Monitor, Re-train
- Hybrid ML
  - Build where you want, deploy where you want
- Governance
  - Ensure that right people get access to the right data

© 2020 IBM Corporation

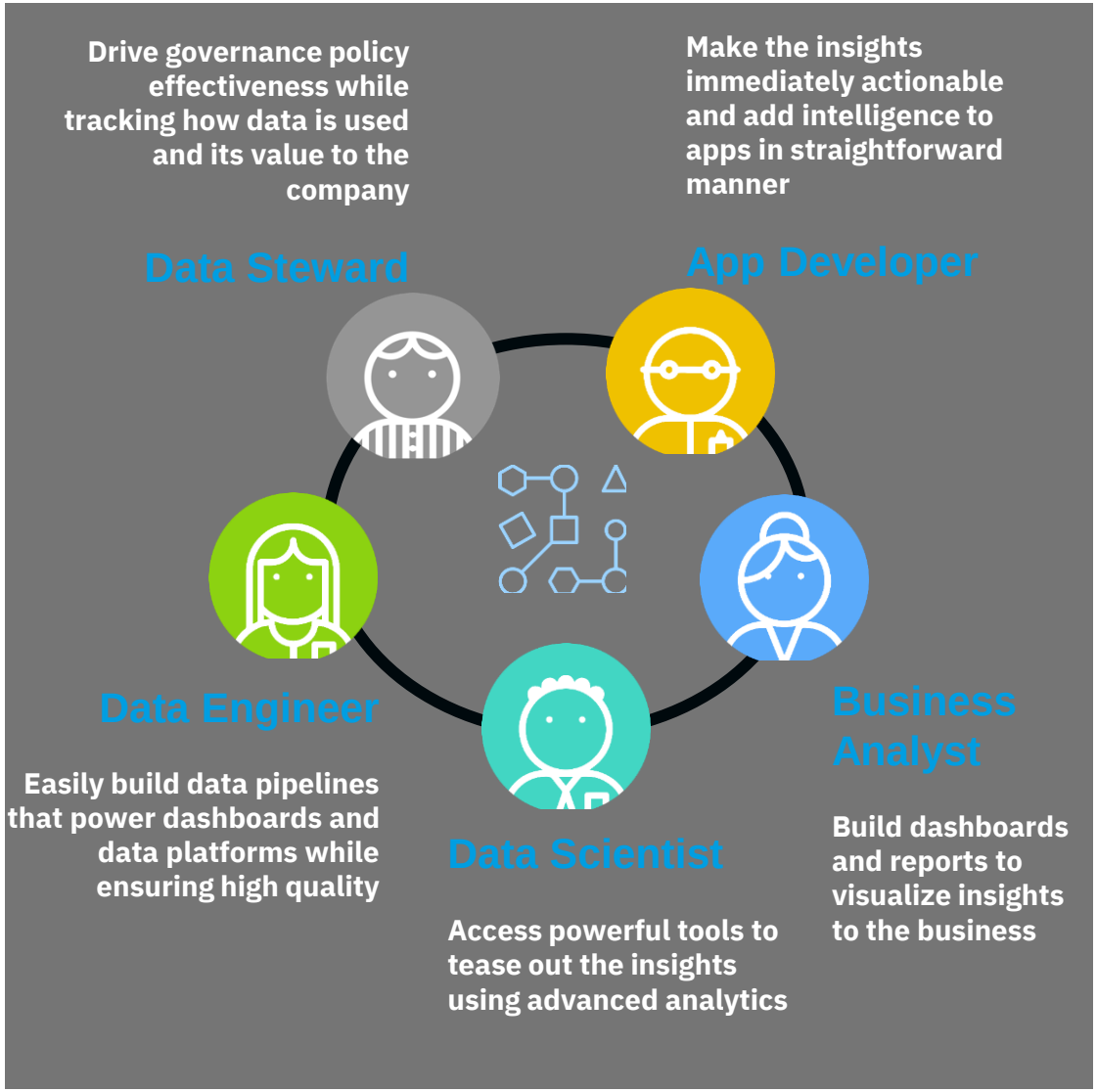
Outline

- Data Science Overview
- Watson Studio Overview
- Lab Overview

© 2020 IBM Corporation

IBM Watson Studio Platform

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



© 2020 IBM Corporation

Watson Studio Deployment Options

- Watson Studio on IBM Cloud
  - Managed offering provided by IBM
- Watson Studio Desktop
- IBM Cloud Private for Data
  - Watson Studio Local

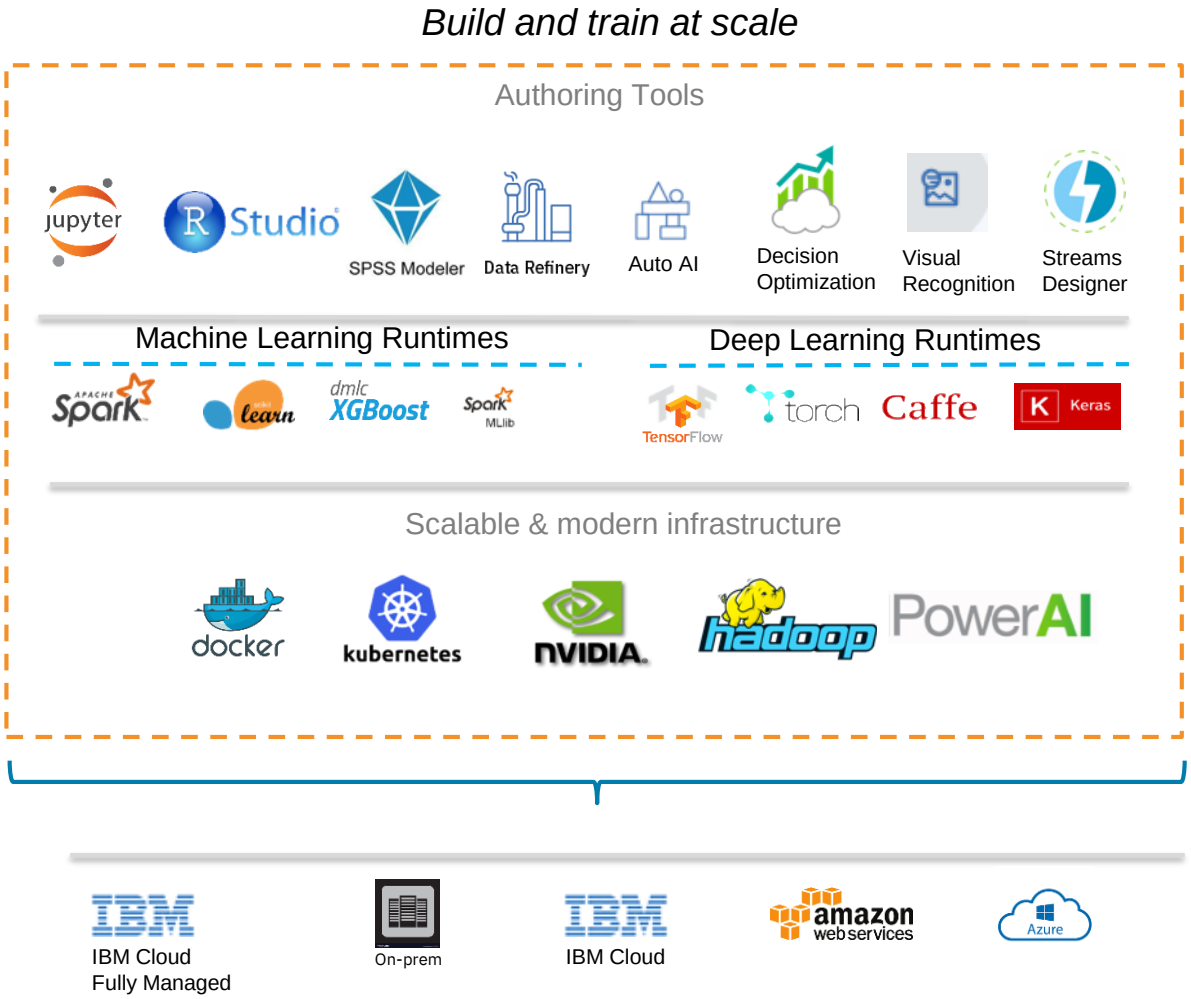
© 2020 IBM Corporation

Watson Studio Tools

- Using best of breed - Open source & IBM tools
- Code (R, Python or Scala) and no-code/visual modeling tools

- Container-based resource management
- Elastic cpu/gpu power
- Run on x86, Power, zLinux
- Integrate with Hadoop/Spark Infrastructure

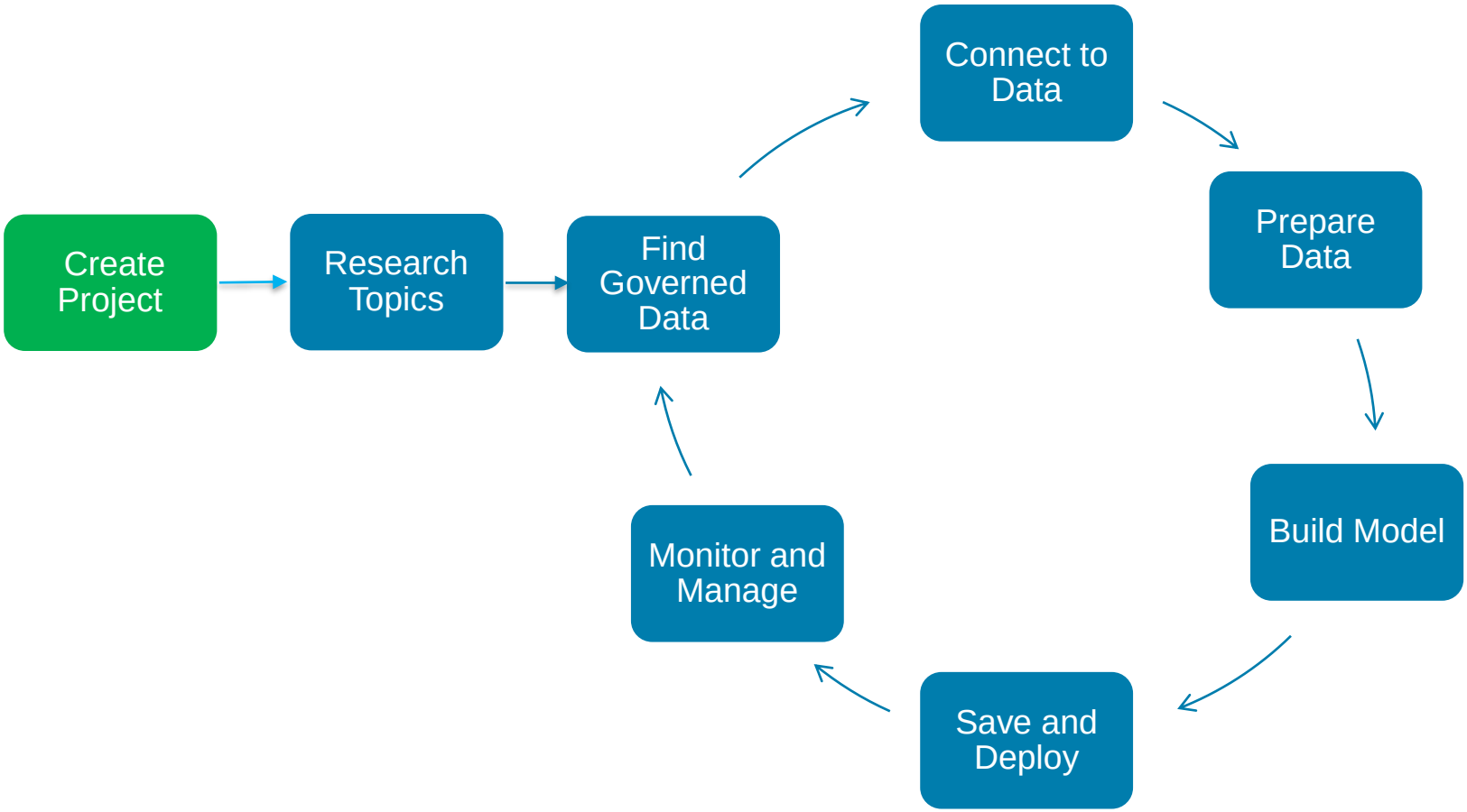
- Train and deploy where your data lives



© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

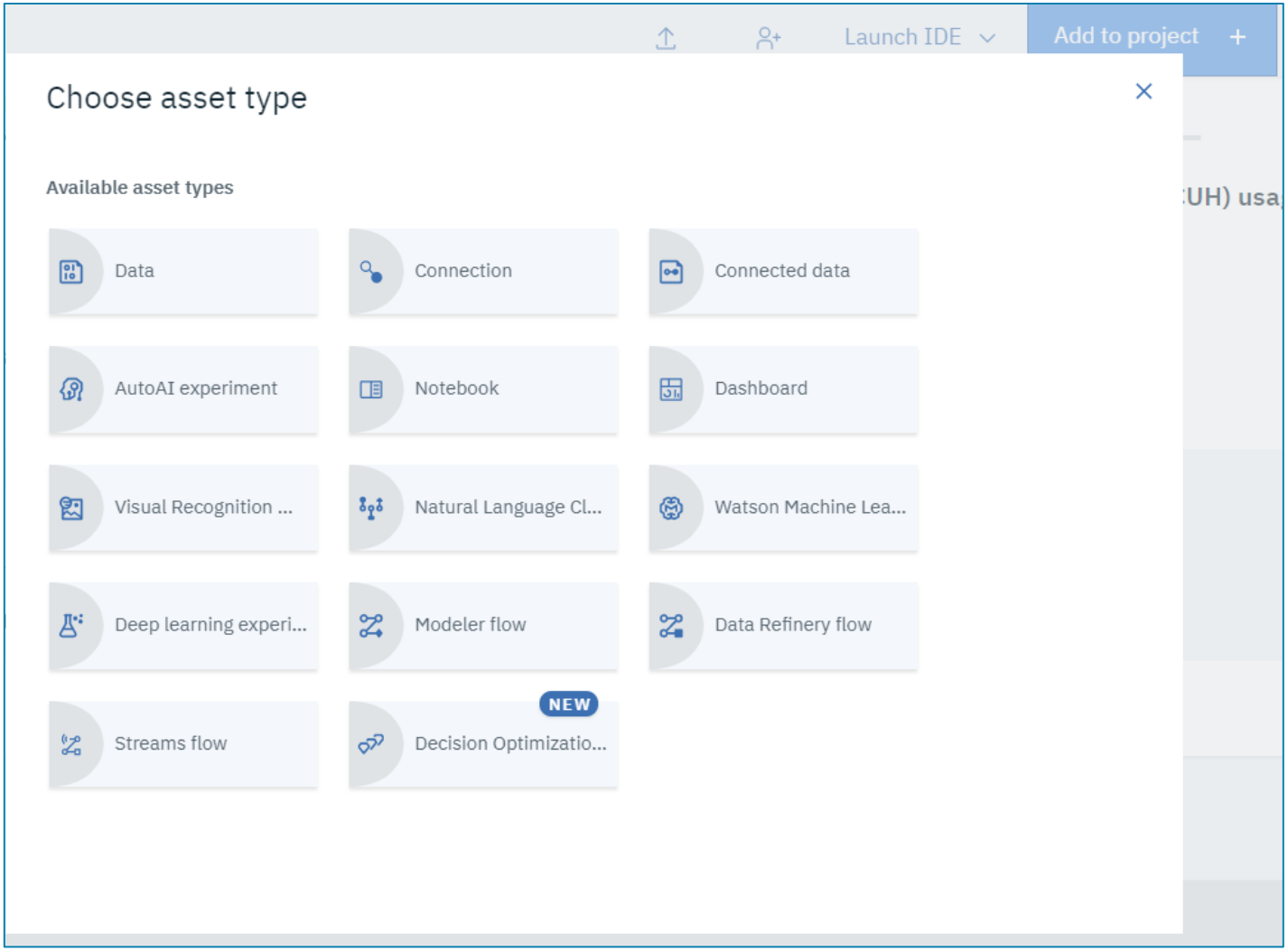
Watson Studio Project Features

Making Data Science a Team Sport

- 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

© 2020 IBM Corporation

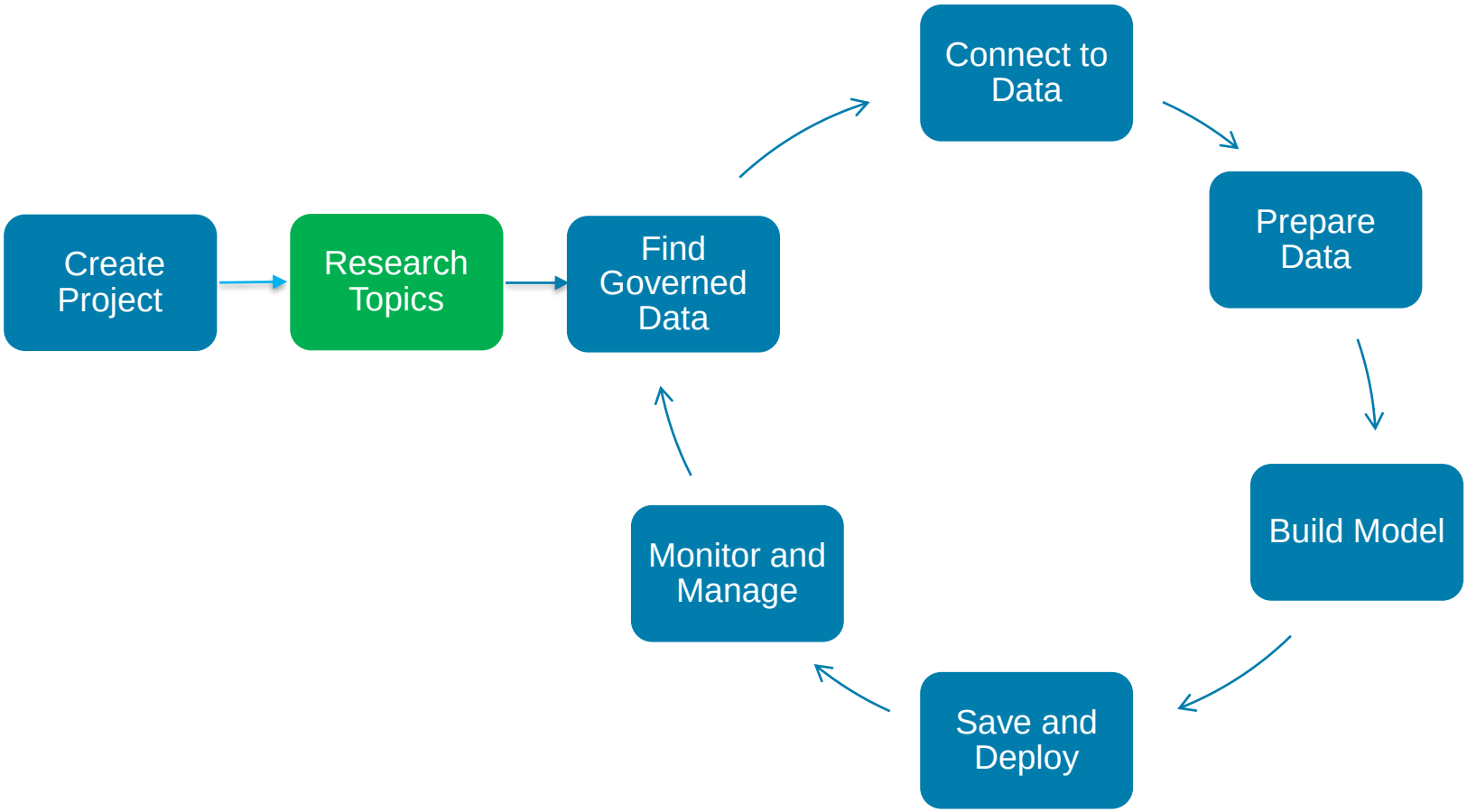
Add to Project



© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

Watson Studio Gallery

Built-in learning to get started

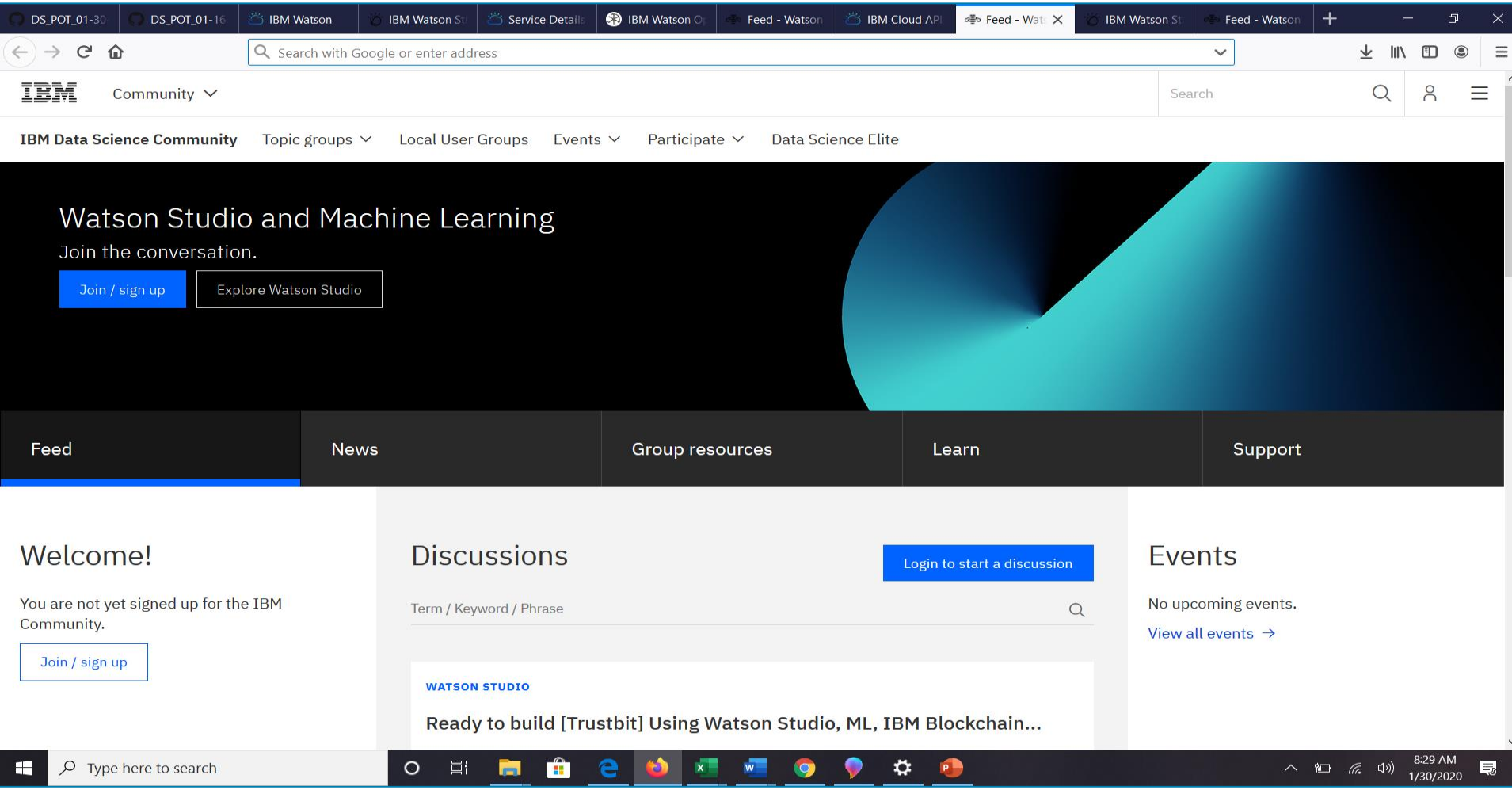
Research Topics

- The Gallery includes notebooks, and data sets
- Copy notebooks or Data Sets into projects
- Continuously updated in IBM's managed service

© 2020 IBM Corporation

Watson Studio Community

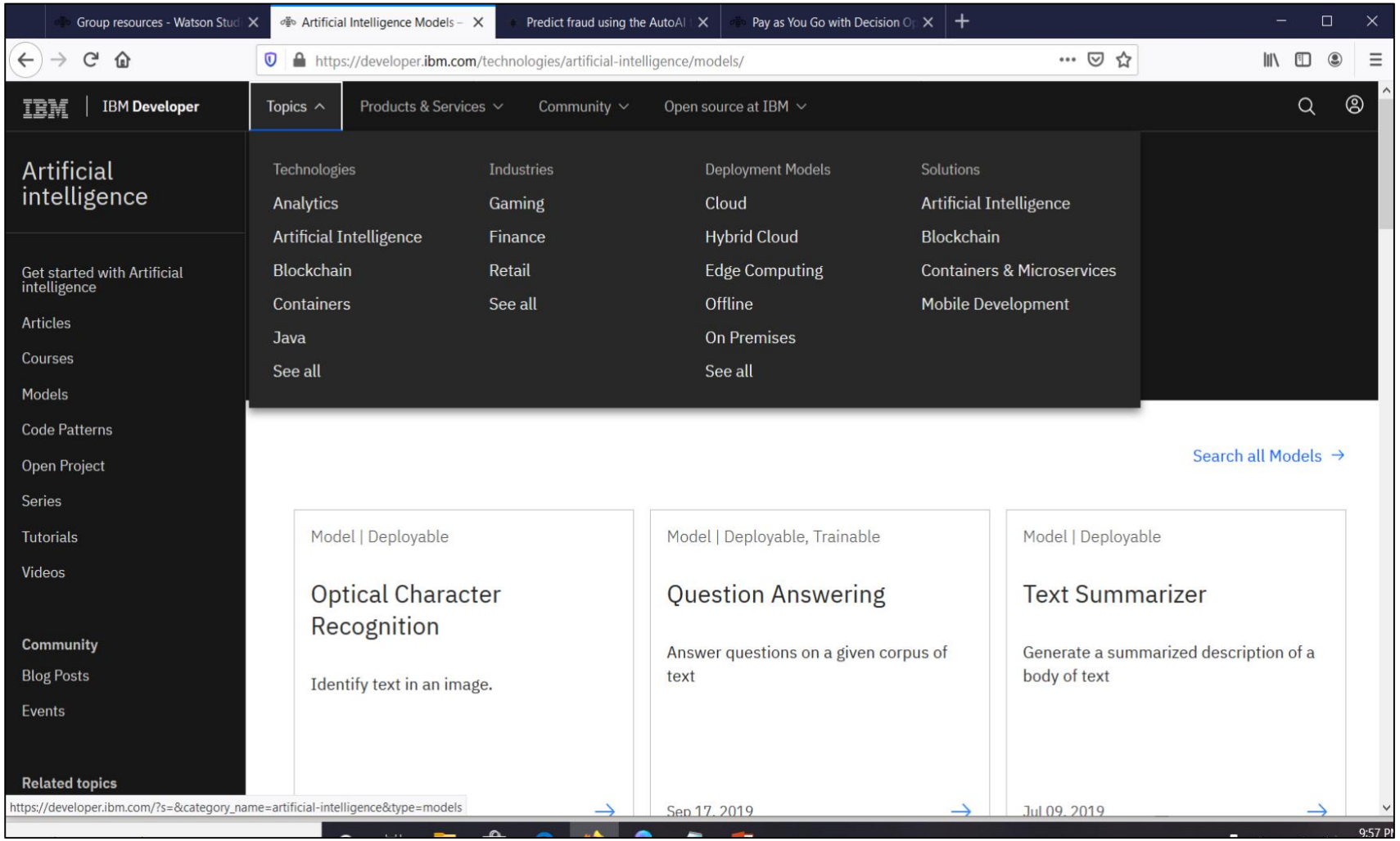
Research Topics



© 2020 IBM Corporation

developer.ibm.com

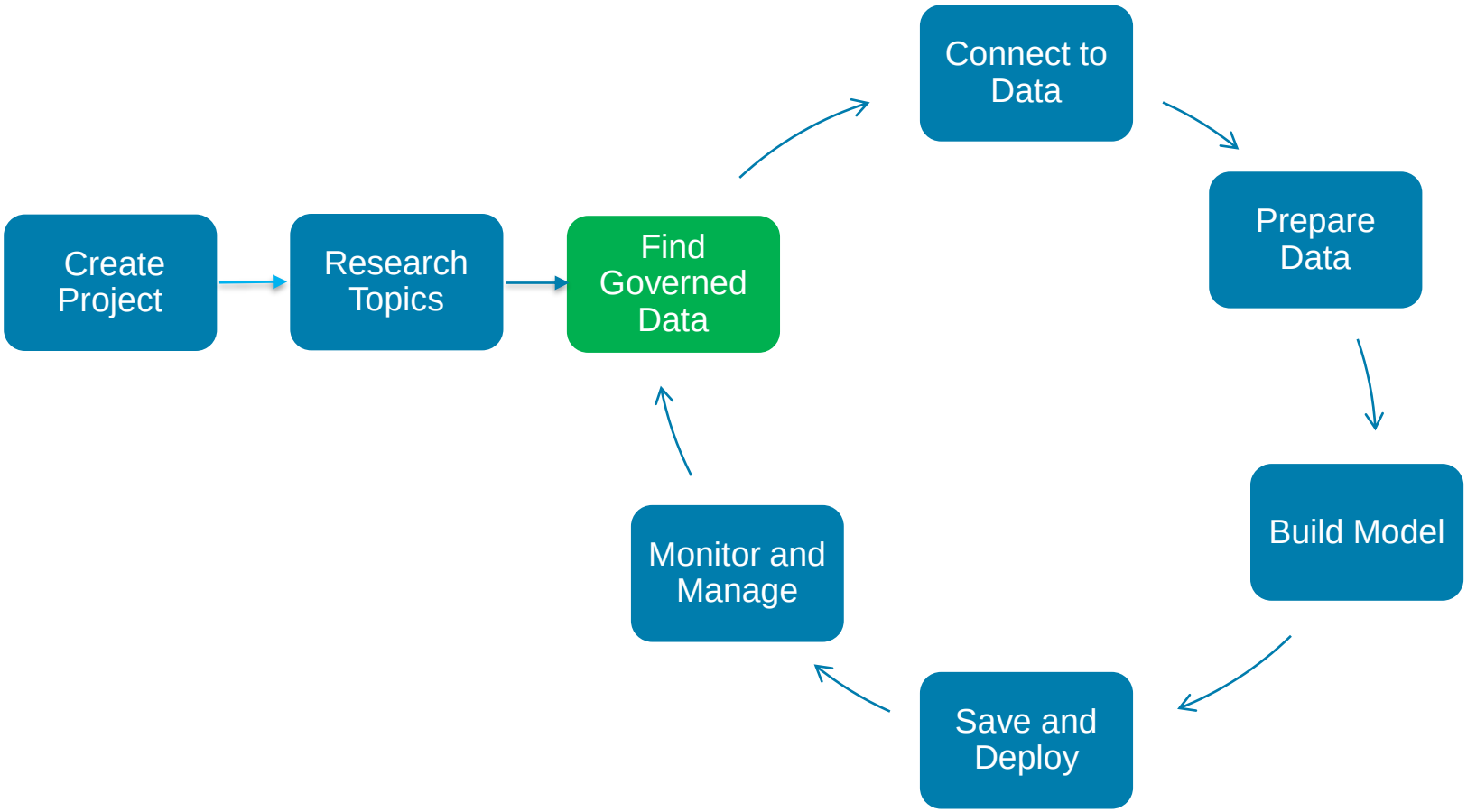
Research Topics



© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

Watson Knowledge Catalog Features

Find Governed Data

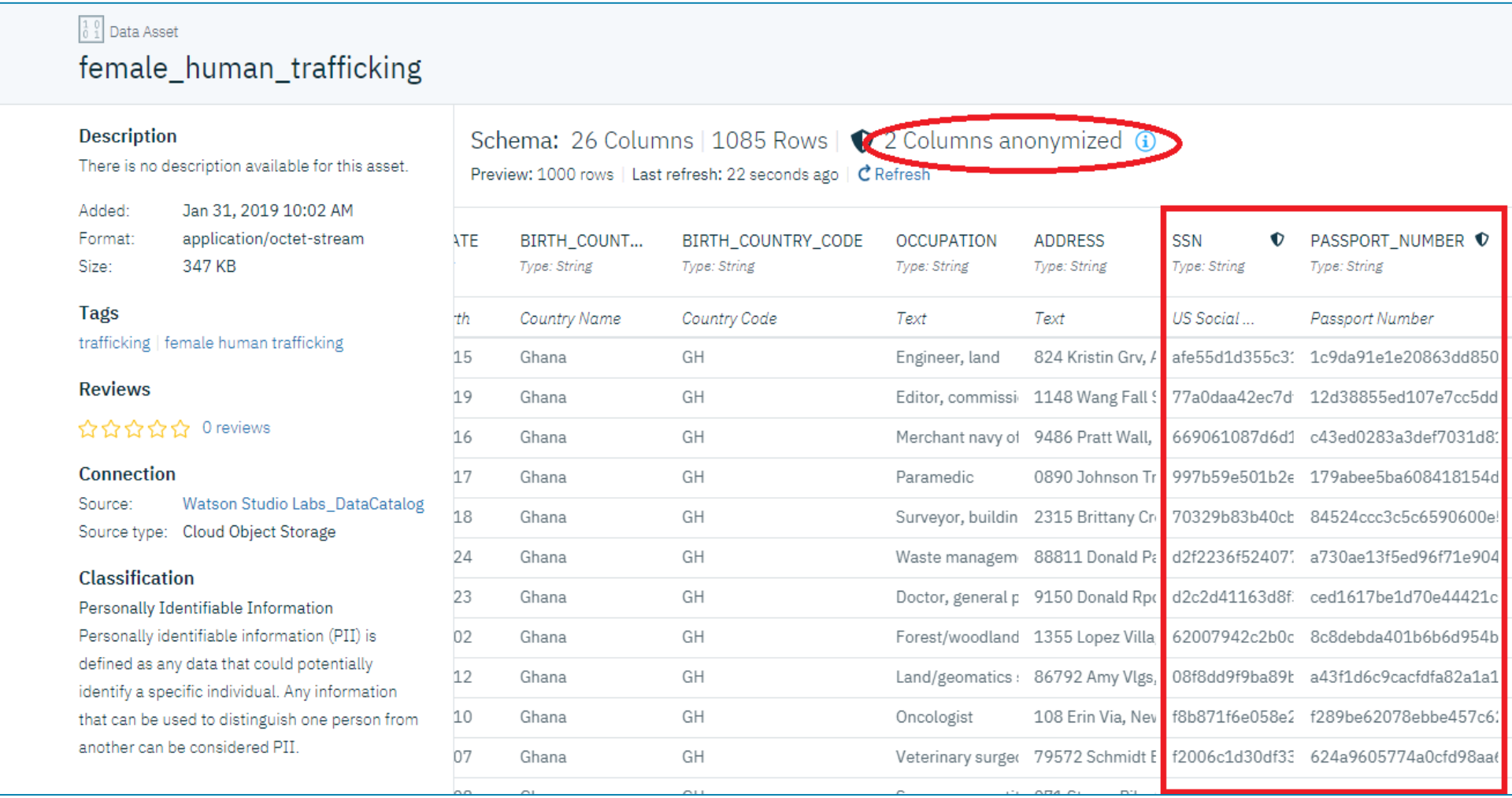
Unlock tribal knowledge and unleash knowledge workers

- **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

© 2020 IBM Corporation

Watson Knowledge Catalog Features

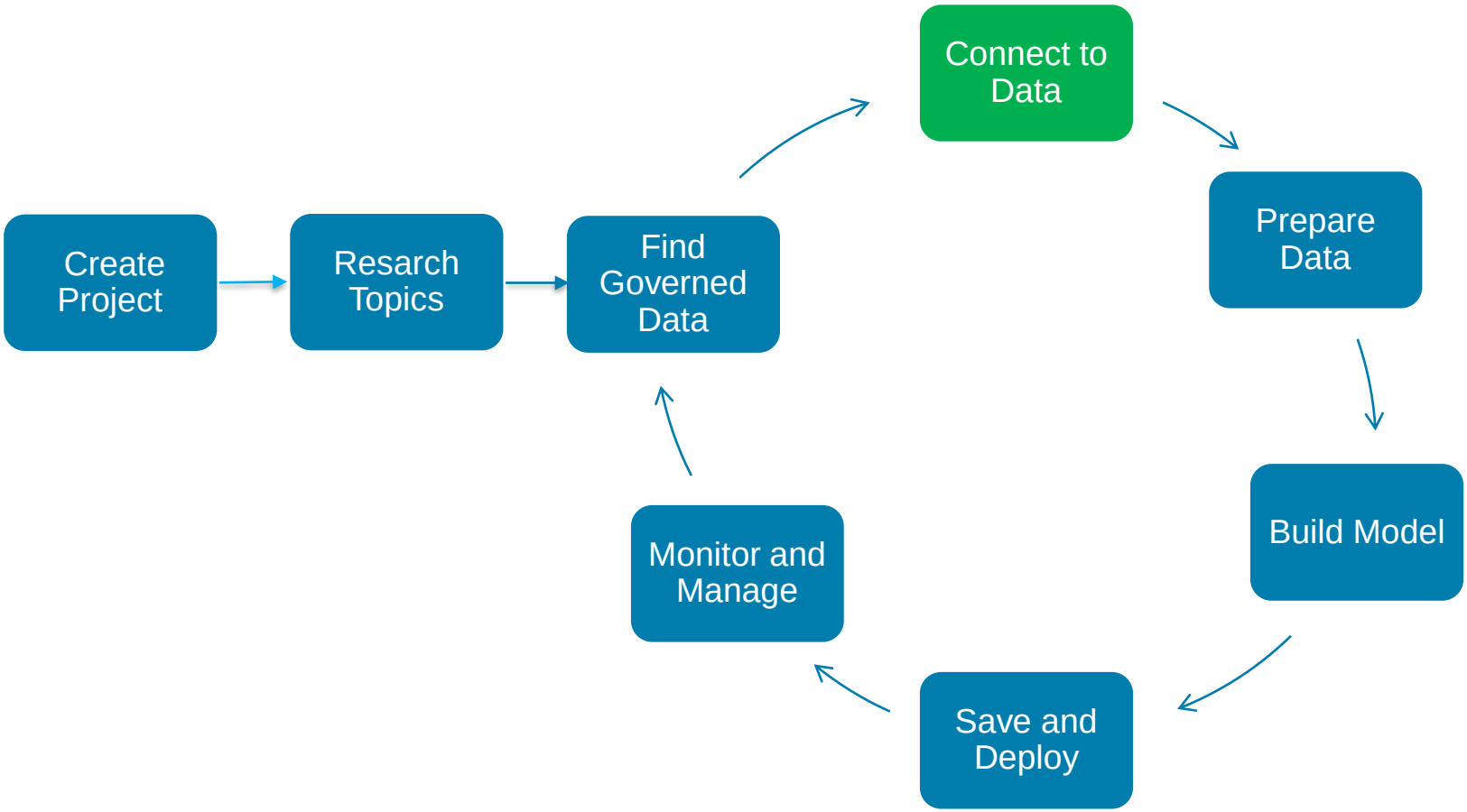
Find Governed Data



© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

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

© 2020 IBM Corporation

Connection Options

Connect to Data

New connection

IBM services

- BigInsights HDFS
- Cloud Object Storage
- Cloud Object Storage (infrastructure)
- Cloudant
- Cognos Analytics
- Compose for MySQL
- Compose for PostgreSQL
- Db2
- Db2 Big SQL
- Db2 for i
- Db2 for z/OS
- Db2 Hosted
- Db2 on Cloud
- Db2 Warehouse
- Informix
- Object Storage OpenStack Swift (infrastructure)
- PureData for Analytics
- Watson Analytics

Third-party services

- Amazon Redshift
- Amazon S3
- Apache Hive
- Cloudera Impala
- Dropbox
- FTP
- Google BigQuery
- Google Cloud Storage
- Hortonworks HDFS
- Looker
- Microsoft Azure Data Lake Store
- Microsoft Azure SQL Database
- Microsoft SQL Server
- MySQL
- Oracle
- Pivotal Greenplum
- PostgreSQL
- Salesforce.com
- Sybase
- Sybase IQ
- Tableau
- Teradata

© 2020 IBM Corporation

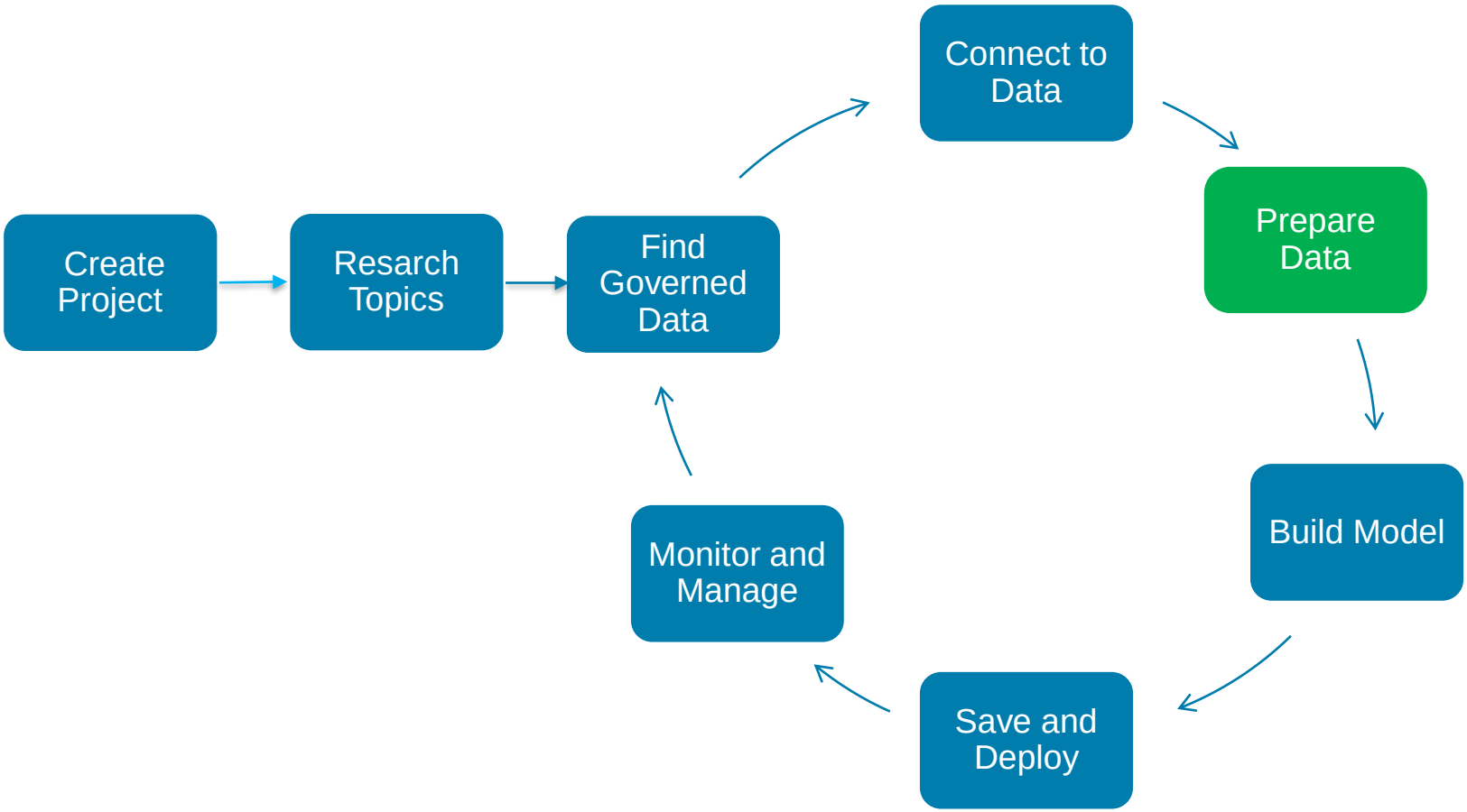
Notebook Screenshot

Connect to Data

© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

Watson Studio Data Refinery Features

Making Data fit for use

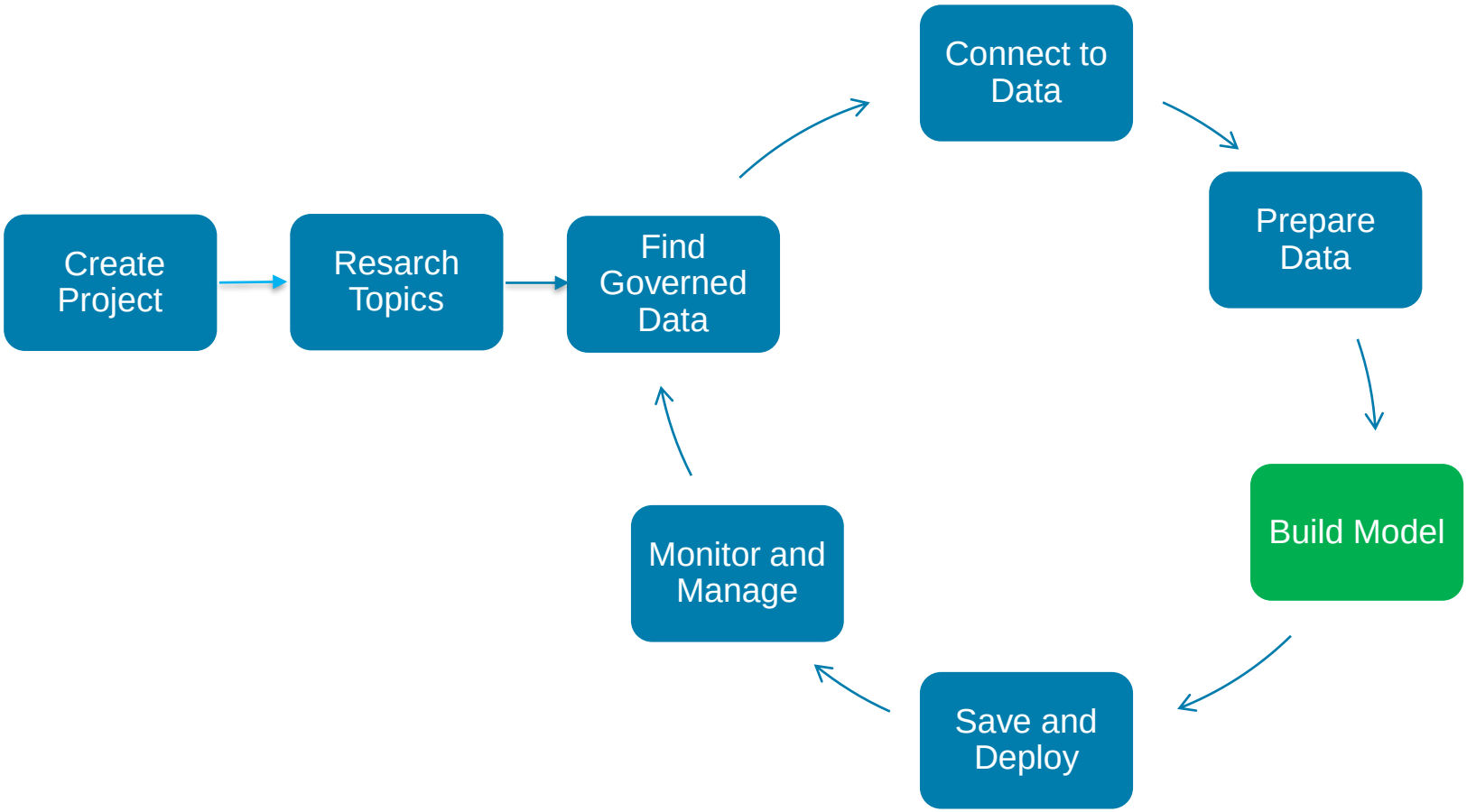
Prepare Data

- 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)

© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

Watson Studio Model Building Features

Build Model

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

Open Source Tools

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

IBM Tools

- AutoAI **
- SPSS Modeler**
- Neural Network Modeler
- Experiment Builder
- Natural Language Classifier Model
- Visual Recognition Model

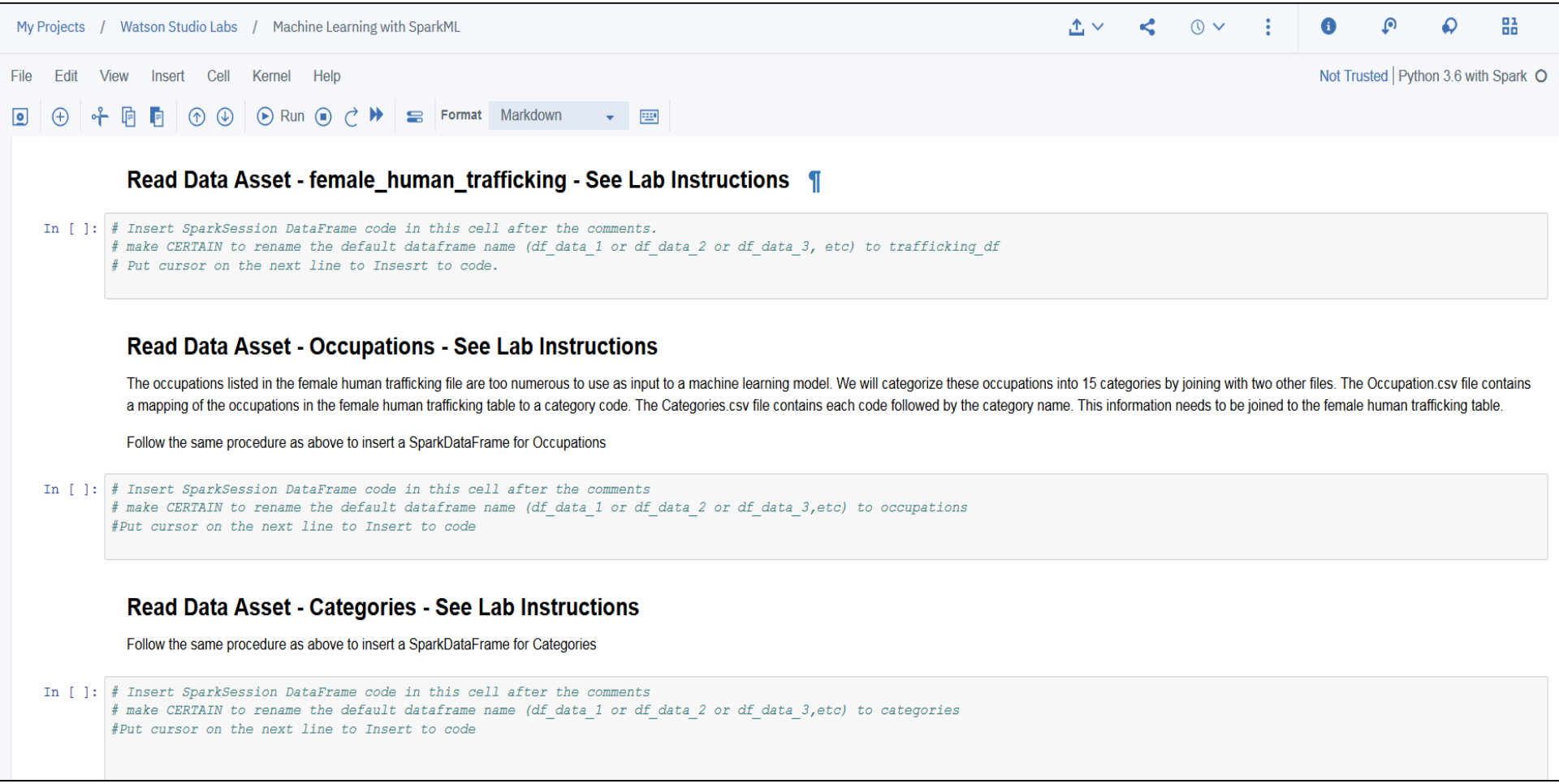
Train at scale on **GPUs** and **distributed** compute

** in hands-on labs

© 2020 IBM Corporation

Jupyter Notebook

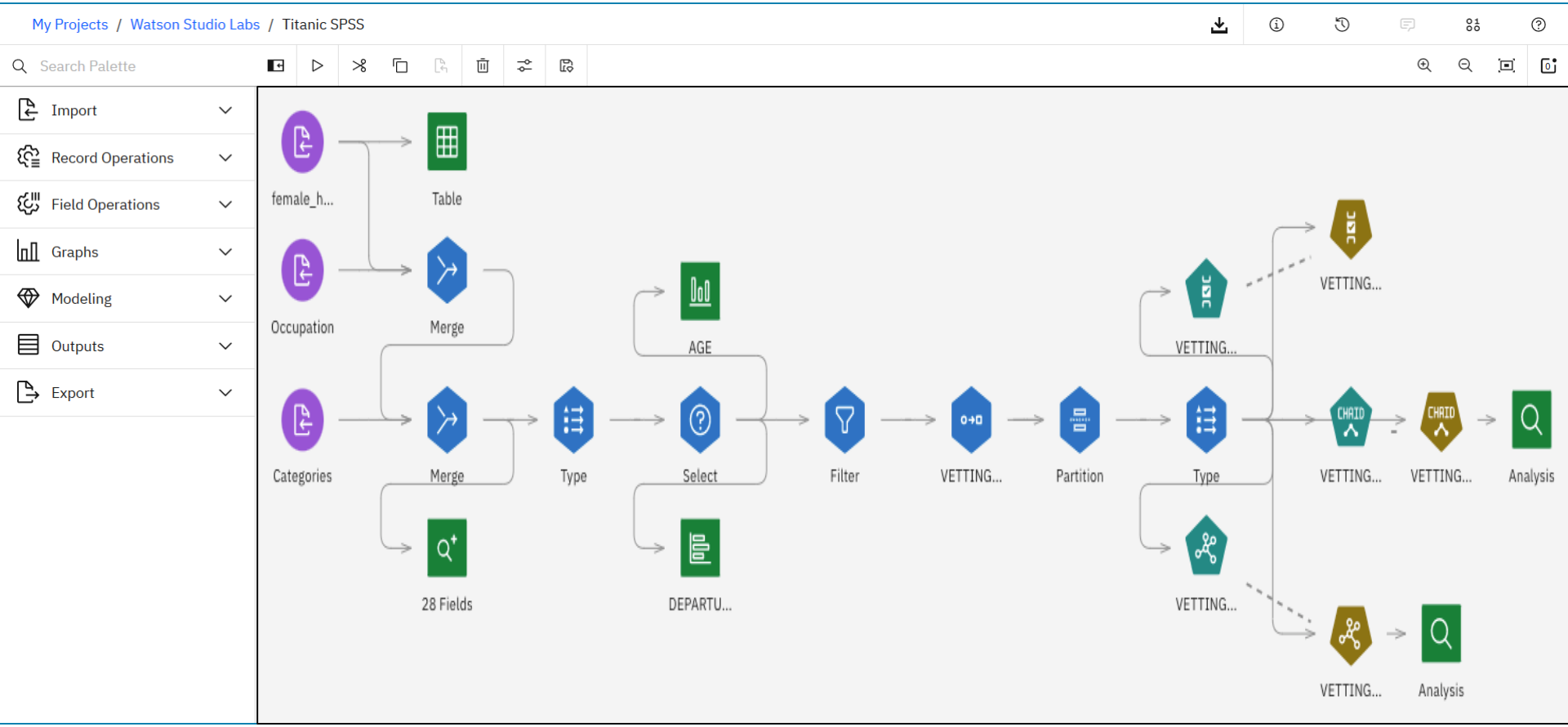
Build Model



© 2020 IBM Corporation

SPSS Modeler

Build Model



© 2020 IBM Corporation

Neural Network Modeler

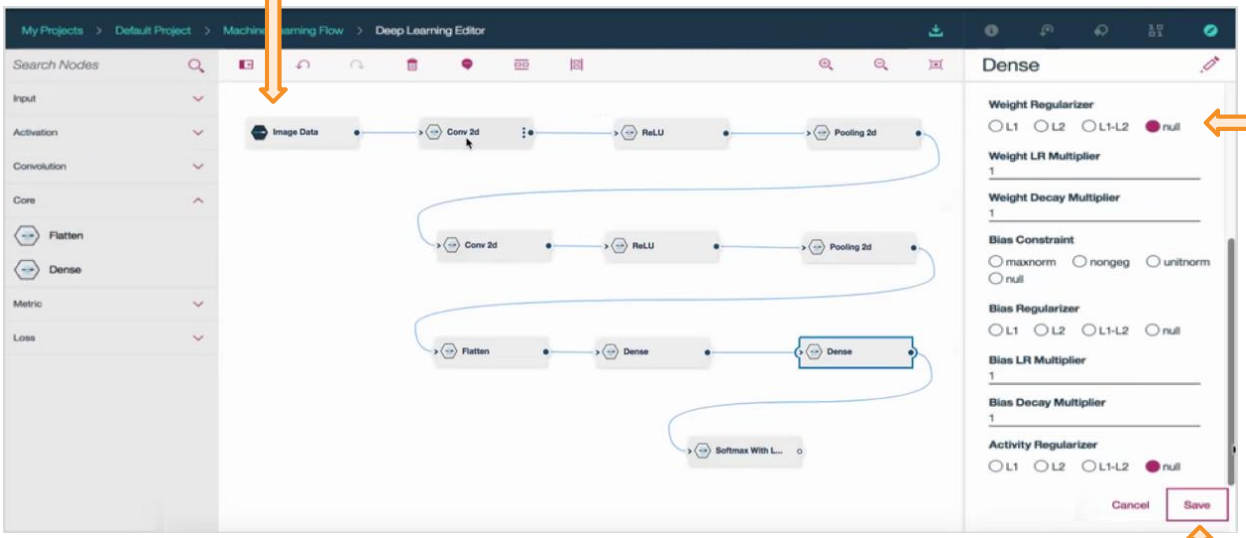
Build Model

An intuitive drag-and-drop, no-code interface for designing neural network structures using the most popular deep learning frameworks. Quickly capture your network design then single click export for experimental optimization.

Supported Frameworks

TensorFlow™ PYTORCH Caffe Keras

Real-time validation of network flow



Drag-and-drop network layers

- Define layer configuration
- Choose optimizer params

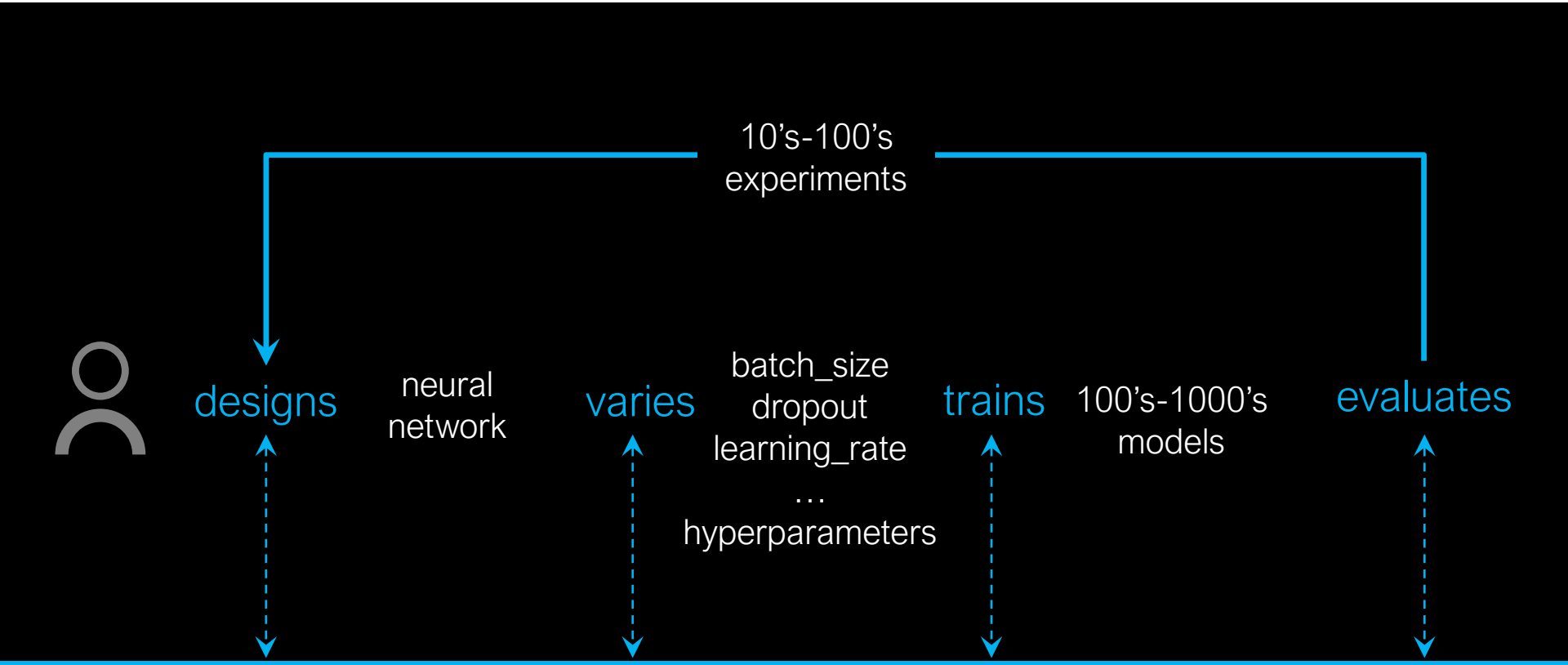
- Generate CPU or GPU compatible code

- Save as popular framework code
- Export as a python notebook
- Execute as batch experiment

© 2020 IBM Corporation

Experiment Builder

Build Model

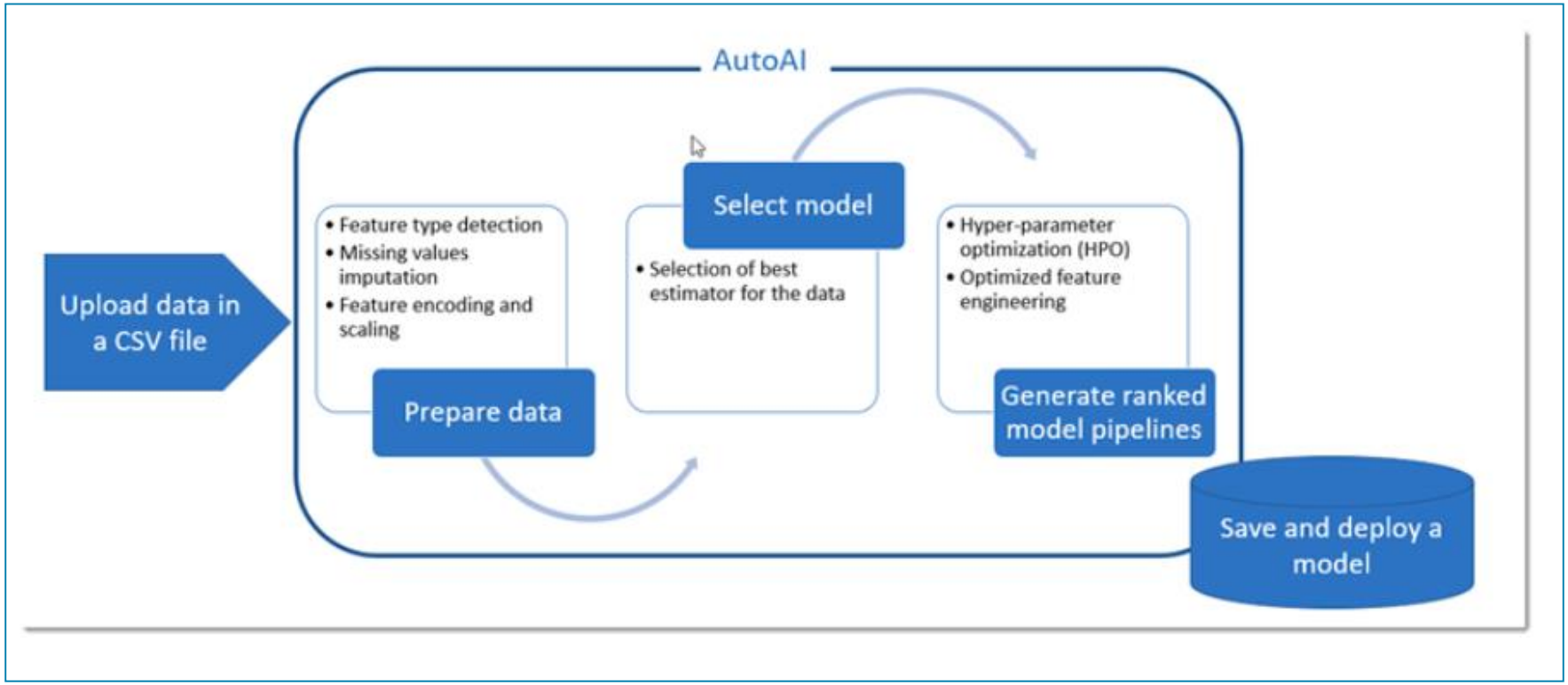


Experiment Builder
supports the end-to-end workflow

© 2020 IBM Corporation

AutoAI

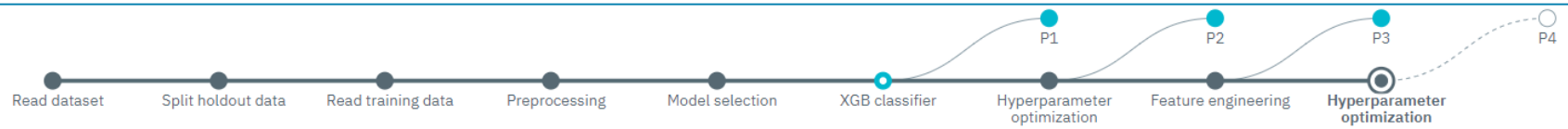
Build Model



© 2020 IBM Corporation

AutoAI

Build Model



Pipeline leaderboard

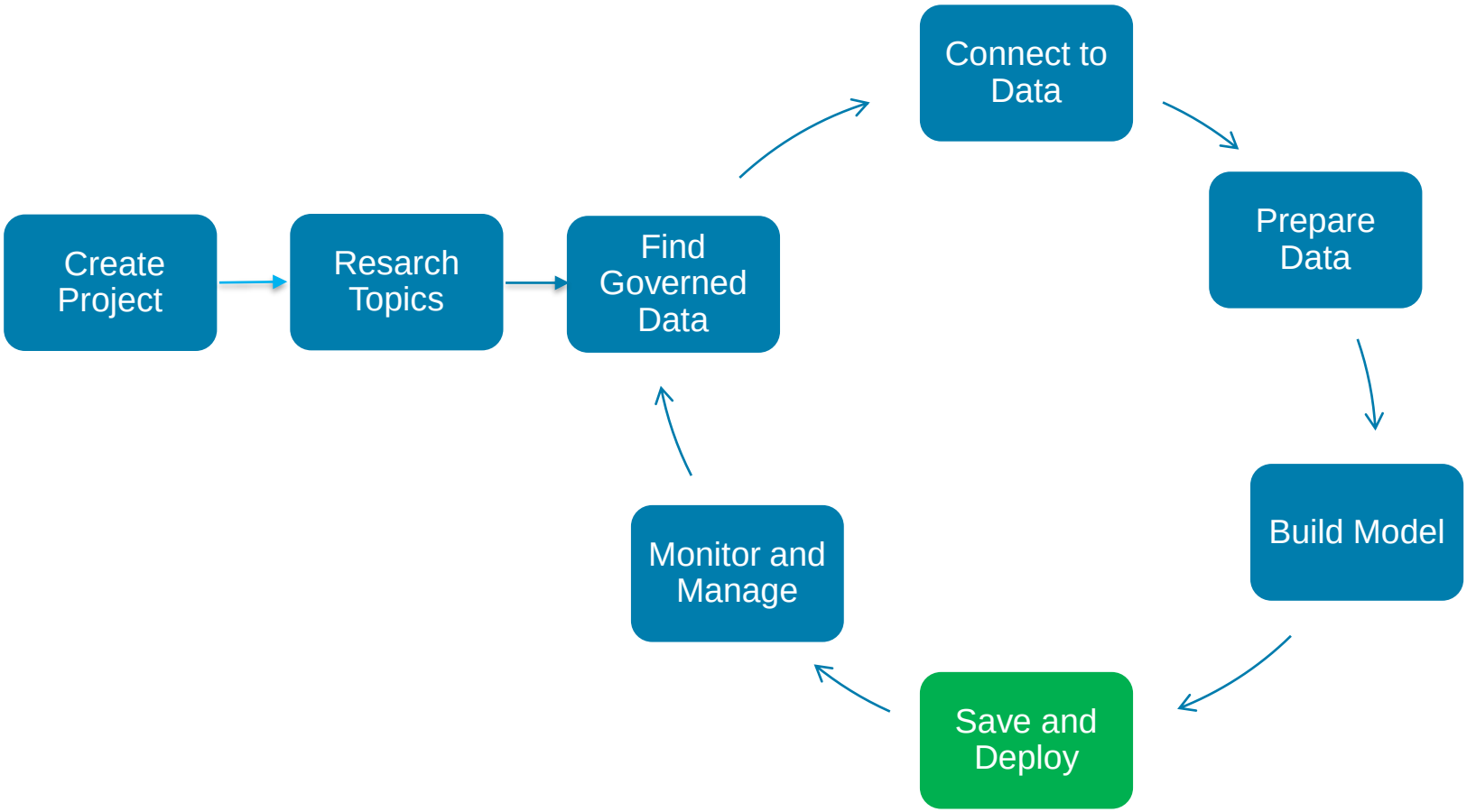
Compare pipelines | Ranking based on: Accuracy

	RANK	ACCURACY	PIPELINE INFORMATION	
>	1	0.897	P3 - XGB classifier estimator Transformers (8): Preprocessing > Standard scaler > Univariate feature selection > Sine > Univariate feature selection > Tangent > ...	View details Save as model
>	2	0.884	P1 - XGB classifier estimator Transformers (2): Preprocessing > XGB classifier estimator	View details Save as model
>	3	0.884	P2 - XGB classifier estimator Transformers (2): Preprocessing > XGB classifier estimator	View details Save as model

© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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

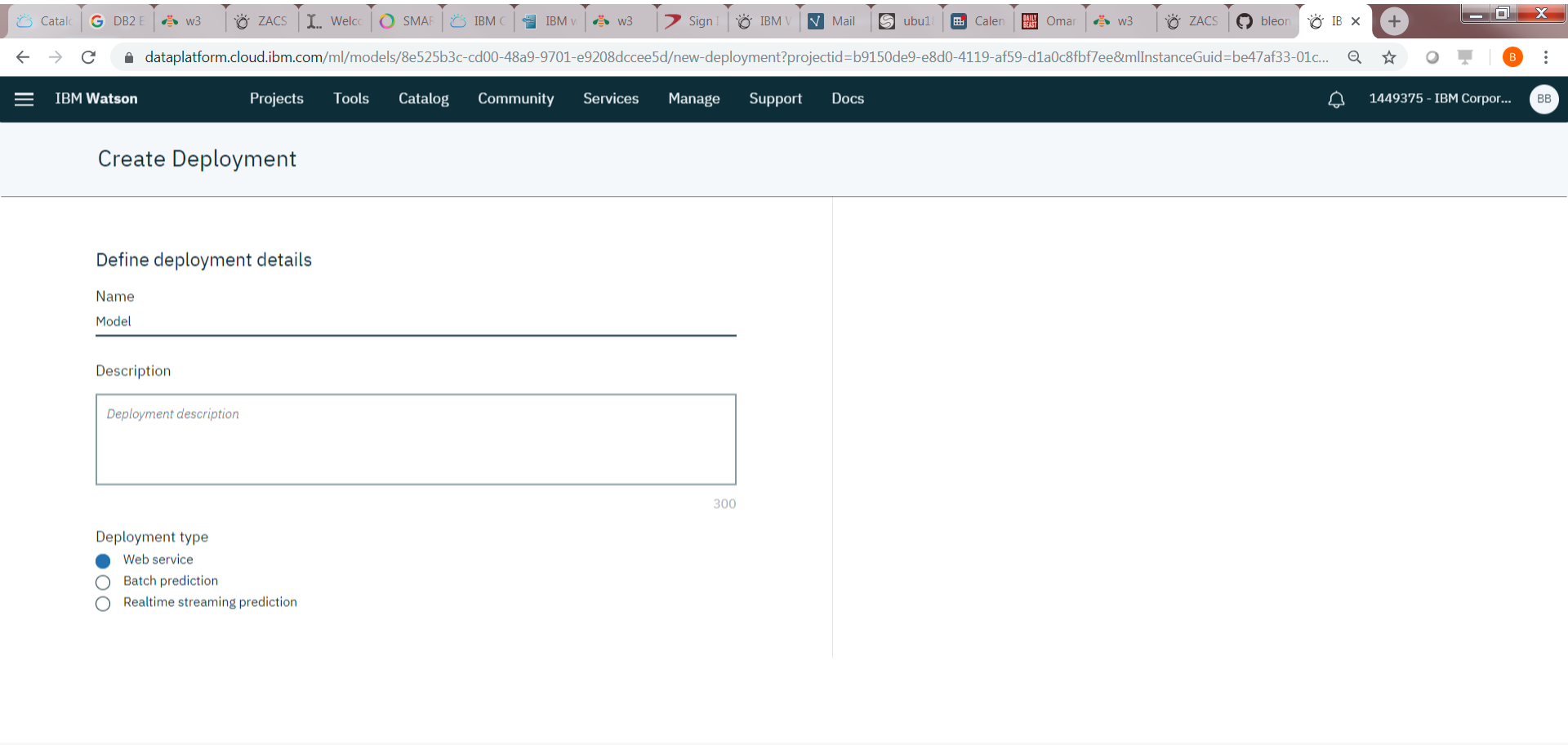


© 2020 IBM Corporation

Watson Studio Save and Deploy Models

Save and Deploy Models with Watson Machine Learning

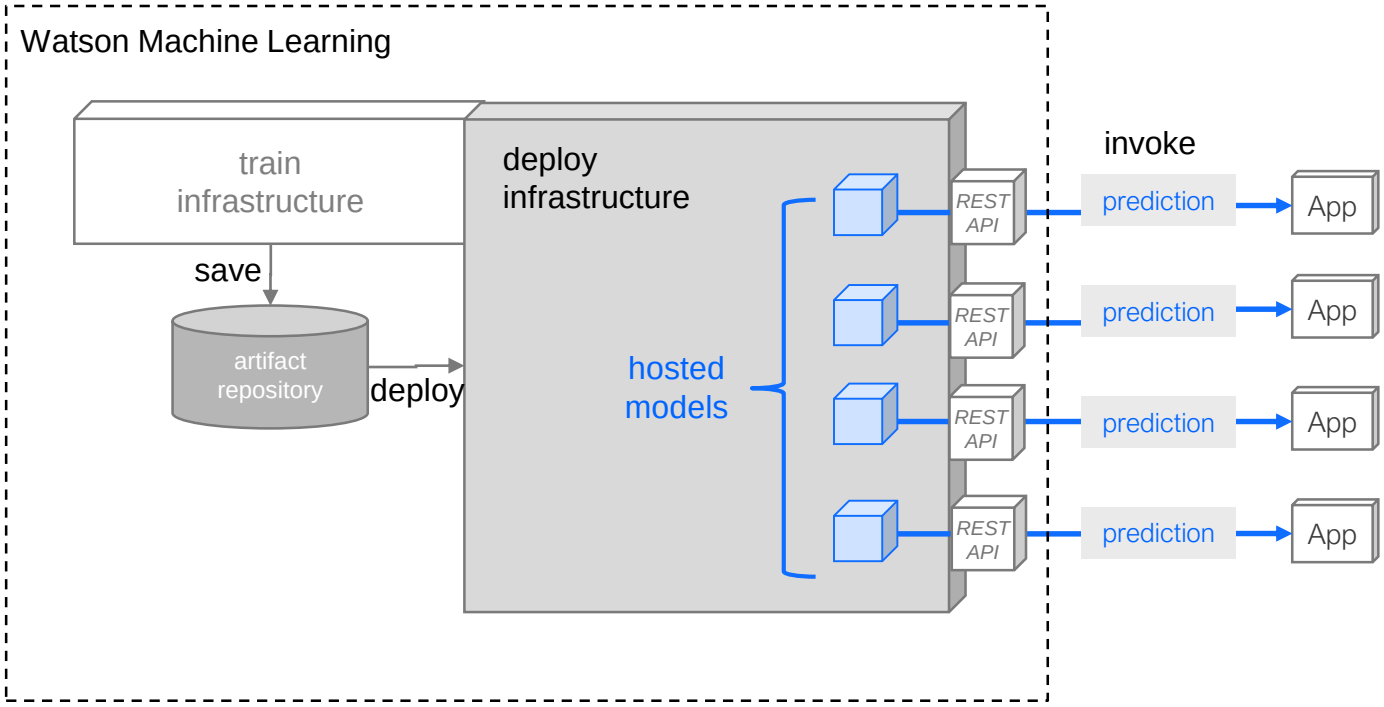
Save and Deploy



© 2020 IBM Corporation

Watson Studio Save and Deploy Trained Models

Save and Deploy Models with Watson Machine Learning



© 2020 IBM Corporation

Watson Studio Save and Deploy Features

Save and Deploy

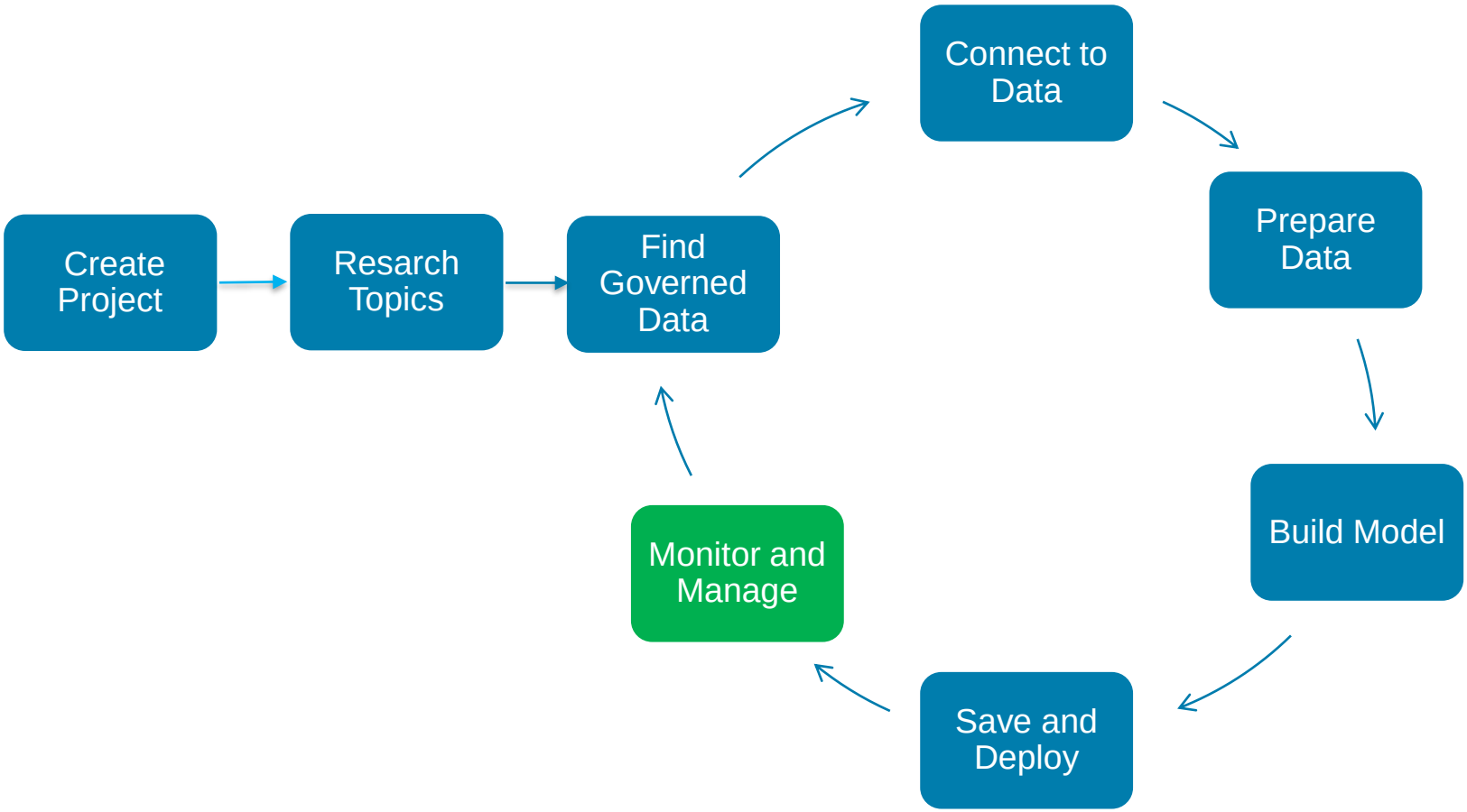
Save and Deploy Models with Watson Machine Learning

- 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

© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2020 IBM Corporation

Our vision for Trusted AI

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

Monitor and Manage

Is it accurate?

Is it fair?

Is it easy to understand ?

Did anyone tamper with it?

© 2020 IBM Corporation

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

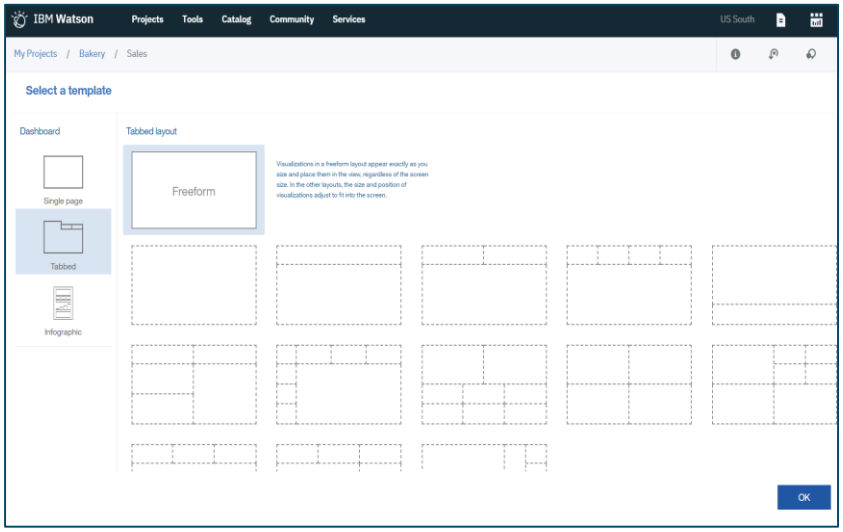
Open By Design

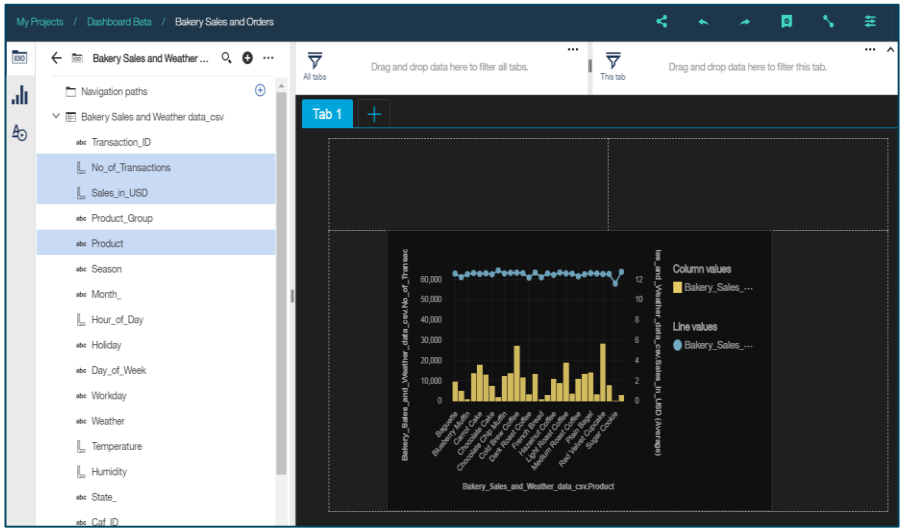
- Monitor models deployed on third party model server engines
- Deploy behind enterprise firewall or on IaaS provider

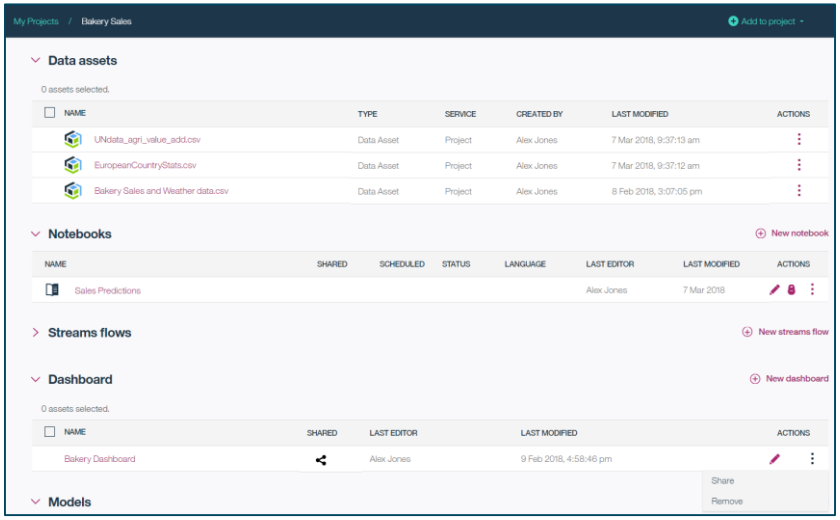
© 2020 IBM Corporation

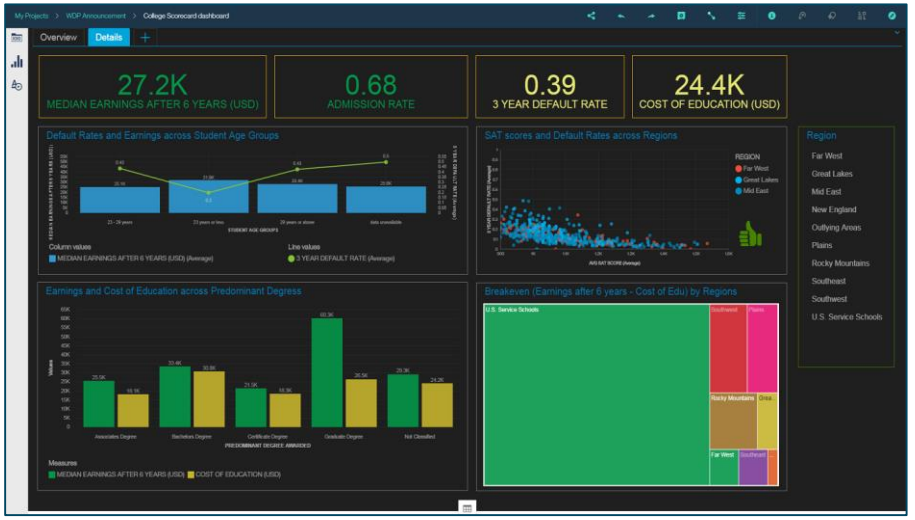
Watson Studio Dynamic Dashboards

Making insights available to all









© 2020 IBM Corporation

Watson Studio Takeaways

Integrated Collaboration Environment

- Data Scientists, Subject Matter experts, Business Analysts & Developers all in one environment to accelerate innovation, collaboration and productivity
- Built-in learning to get started or go the distance with advanced tutorials

Choice of Tools for the full AI lifecycle

- Best in-breed open source and IBM tools that support the end-to-end AI lifecycle
- Choice of code or no-code tools to build and train your own ML/DL models or easily train and customize pre-trained Watson APIs

Support for all levels of expertise

- Use Watson smarts and recommendations for the best algorithms to use given your data, OR
- Use the rich capabilities and controls to fine tune your models

Multiple Deployment Options

- Watson Studio on IBM Cloud – Managed offering
- Watson Studio Local – Private Cloud, Public Cloud-(IBM, Azure, AWS)
- Watson Studio Desktop

Model lifecycle & management

- Deploy models into production then monitor them to evaluate performance.
- Capture new data for continuous learning and retrain models so they continually adapt to changing conditions.

Integrated with Knowledge Catalog

- Intelligent discovery of data and AI assets that enables reuse & improves productivity
- Seamlessly integrated for productive use with Machine Learning and Data science
- Powerful governance tools to control and protect access to data

46

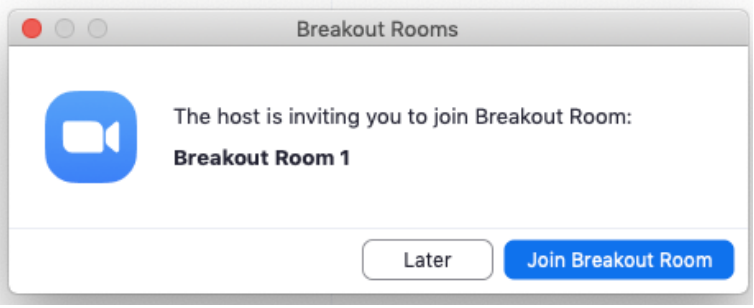
© 2020 IBM Corporation

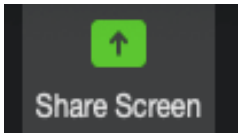
Outline

- Data Science Overview
- Watson Studio Overview
- Lab Overview

© 2020 IBM Corporation

Breakout Rooms





© 2020 IBM Corporation

Lab Use Case: Female Human Trafficking

Input

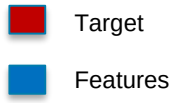
- Generated fake travel records based on incoming custom forms.
- Subset of records were vetted as “high”, “medium”, or “low” risk for Female Human Trafficking by an analyst.

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.

© 2020 IBM Corporation

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.



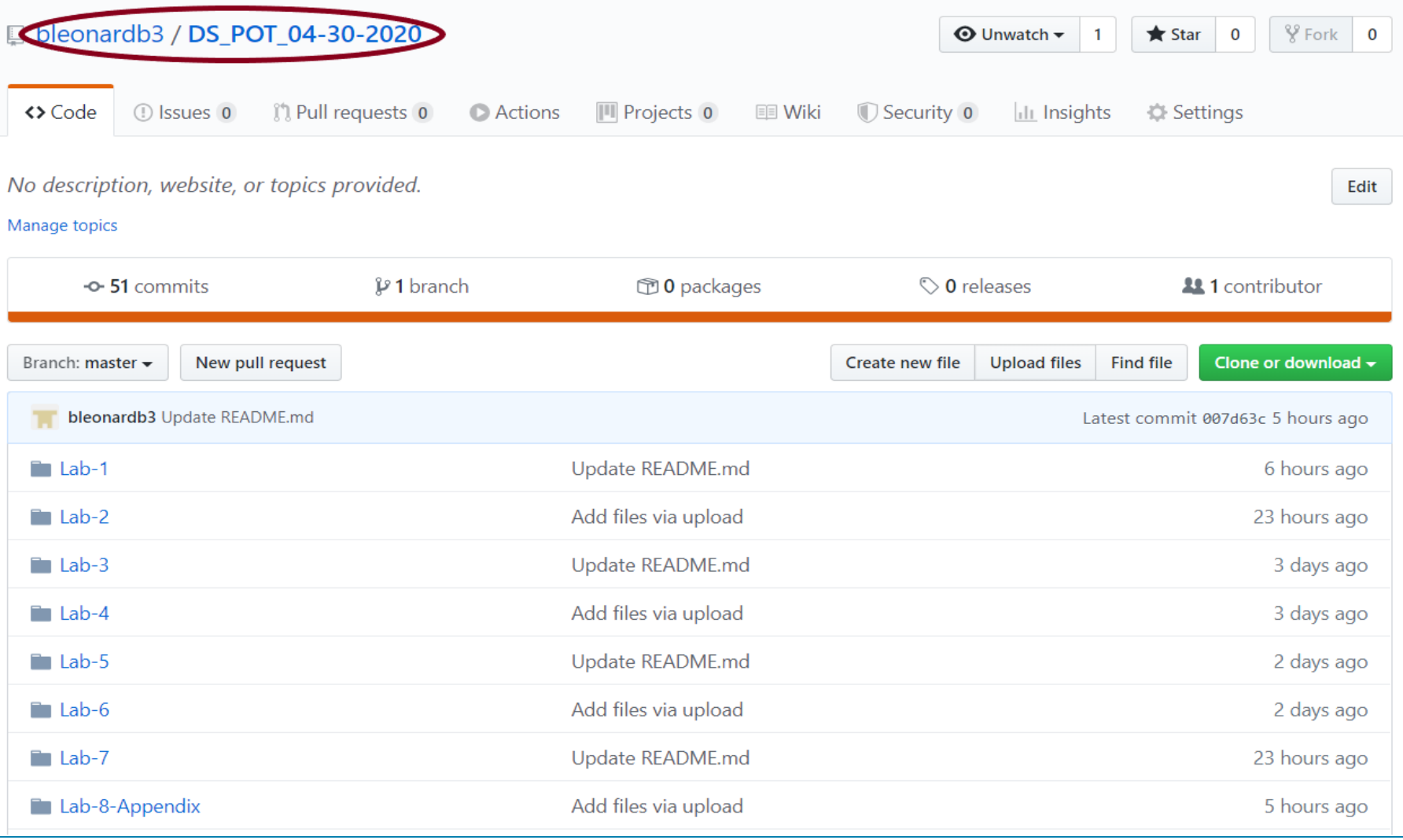
© 2020 IBM Corporation

Lab Tips

- Watson Studio url: dataplatform.ibm.com
- Labs are in www.github.com/bleonardb3/DS_POT_04-30-2020 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.
- All of the Labs should be done in the project that you created in Lab-1

© 2020 IBM Corporation

Github Repository



© 2020 IBM Corporation

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.

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-Appendix - This lab will fulfill the prerequisites for Labs 3,4, and 5, if Lab-2 is not completed successfully.

© 2020 IBM Corporation

Github Repository

Lab-1 Readme

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 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

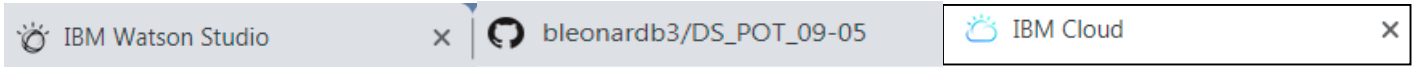
Instructions:

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

Instructions.

© 2020 IBM Corporation

Browser Tabs



© 2020 IBM Corporation

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

© 2020 IBM Corporation

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.

© 2020 IBM Corporation

End-to-End Data Science using Watson Studio

Lunch Break. We will return for review at 12:30 pm.

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.

© 2020 IBM Corporation

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.

© 2020 IBM Corporation

End-to-End Data Science using Watson Studio

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

Categories of Machine Learning

Supervised learning

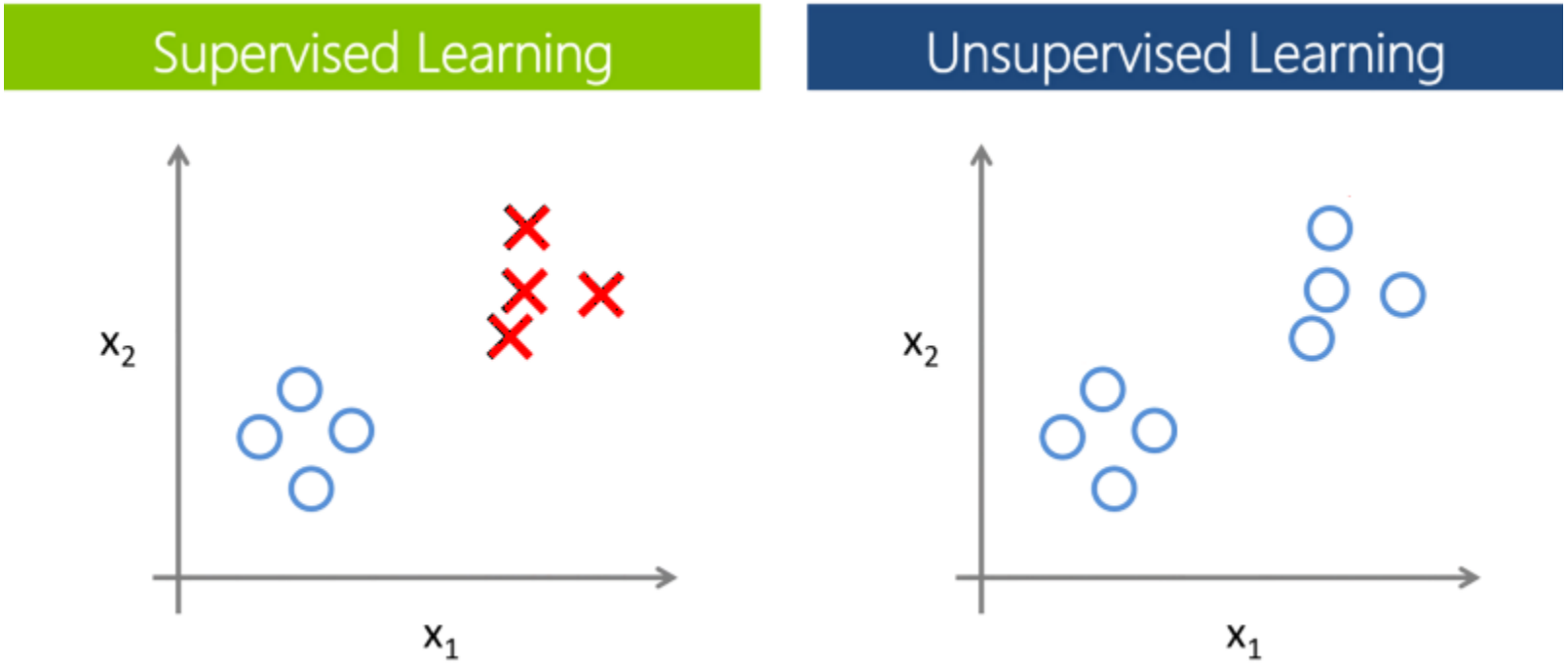
- The program is “trained” on a pre-defined set of “training examples”, which then facilitate its ability to reach an accurate conclusion when given new data
- The algorithm is presented with example inputs and their outcomes (labels)
- The goal is to learn a general rule that maps inputs to outputs

Unsupervised learning

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

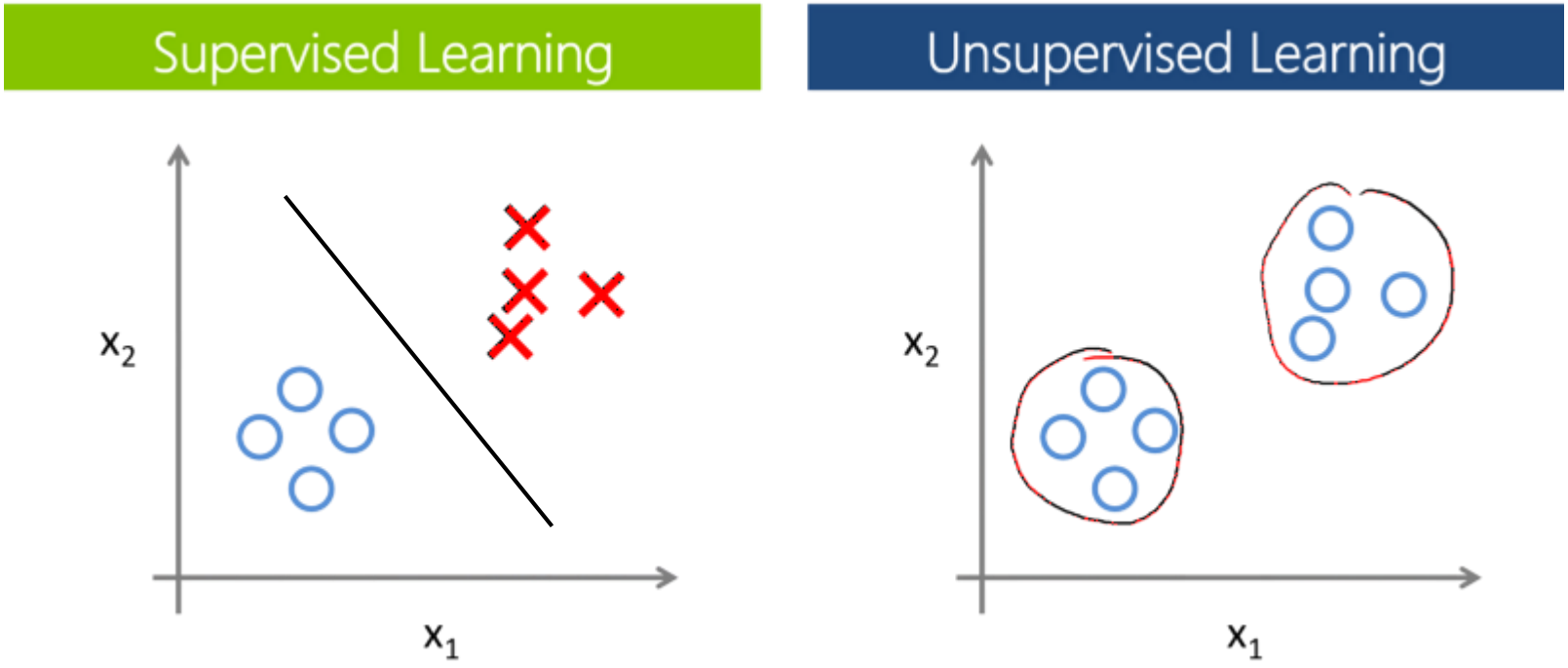
© 2019 IBM Corporation

Supervised vs. Unsupervised Learning



© 2019 IBM Corporation

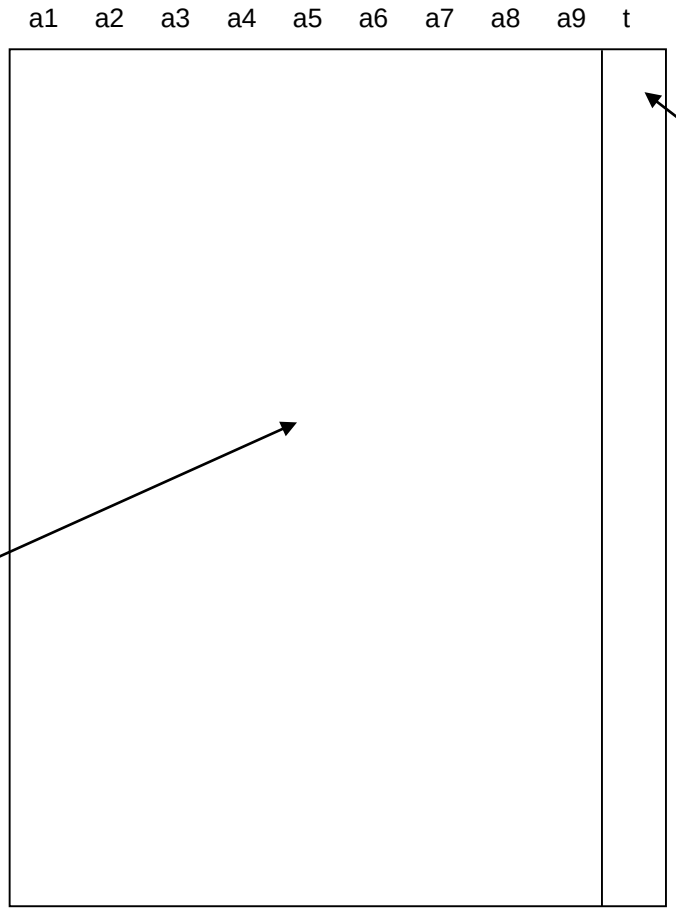
Supervised vs. Unsupervised Learning



© 2019 IBM Corporation

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

© 2019 IBM Corporation

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

© 2019 IBM Corporation

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**

© 2019 IBM Corporation

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

© 2020 IBM Corporation

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

© 2020 IBM Corporation

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)
- Write results into the Database

Save the model in the Model Repository

- Model properties can be saved as well (e.g Area under the ROC curve)

© 2020 IBM Corporation

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 and load into a new DB2 table.
- Save the trained model

© 2020 IBM Corporation

Lab-6: AutoAI + DevOps

Introduction:

This lab consists of two parts. The first part will demonstrate the exciting AutoAI capability to build and deploy an optimized model based on the trafficking data set. The second part will deploy a web application using the IBM Cloud DevOps toolchain that will invoke the deployed model to predict the trafficking risk.

Objectives:

Upon completing the lab, you will:

- Become familiar with the AutoAI feature of Watson Studio.
- Train/Evaluate a machine learning model
- Deploy a machine learning model.

- Deploy a Python Flask web application that we will configure to "call" the deployed machine learning model.
- Configure the application to connect to the machine learning service.
- Update the code in the application to specify the endpoint of the deployed model, and use DevOps to build and re-deploy the application.
- Run the application to demonstrate the use of the deployed machine learning model to score the trafficking risk of a passenger.

© 2020 IBM Corporation

End-to-End Data Science using Watson Studio

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

Trusted AI

Is it accurate?

Is it fair?

Is it easy to understand ?

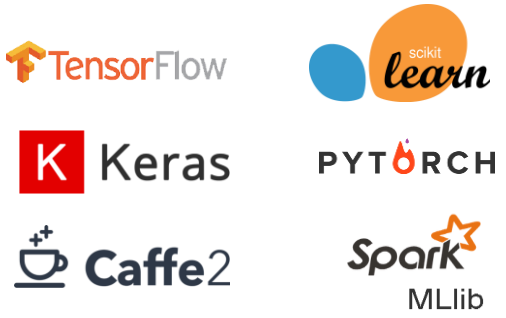
Did anyone tamper with it?

© 2020 IBM Corporation

Watson OpenScale

- **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
  - Provides 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 mode server engines
  - Deploy behind enterprise firewall or on IaaS provider.

Model build / train frameworks



Model serving environments

© 2020 IBM Corporation

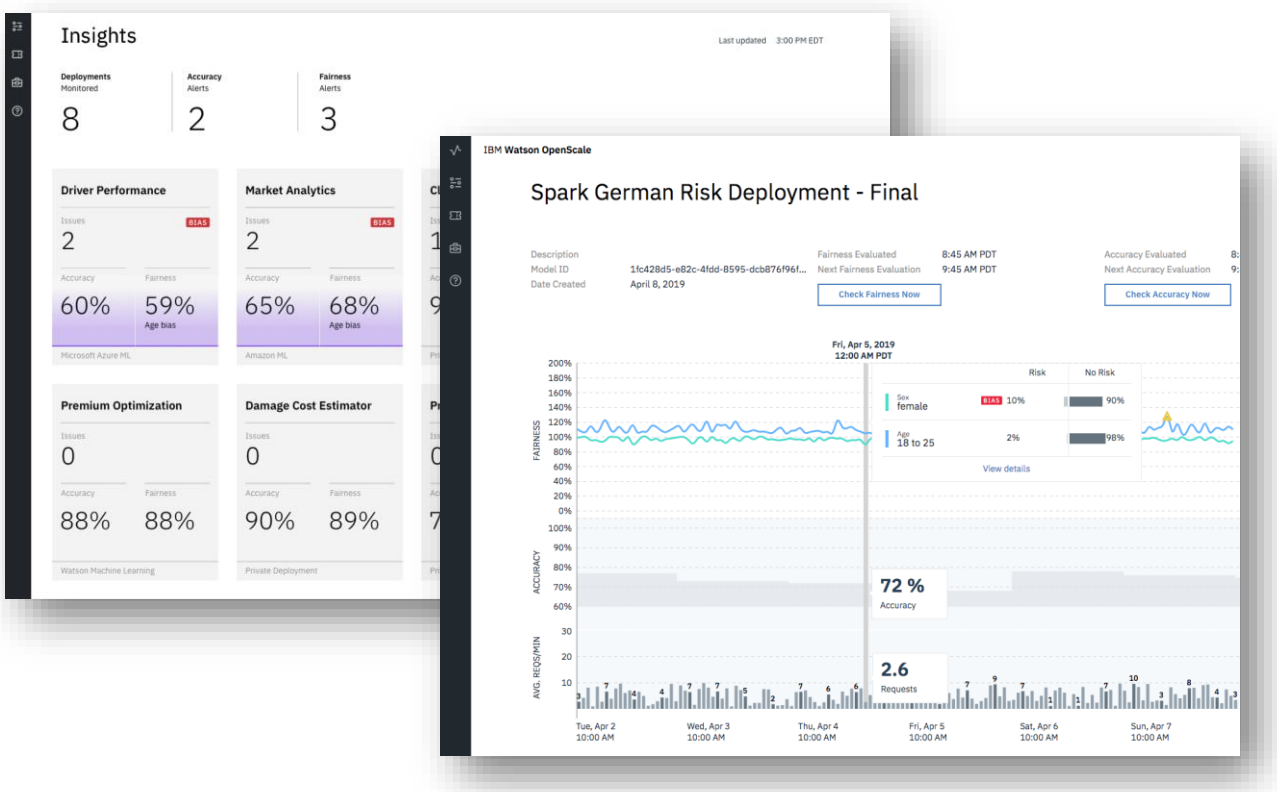
Operations dashboard

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



© 2020 IBM Corporation

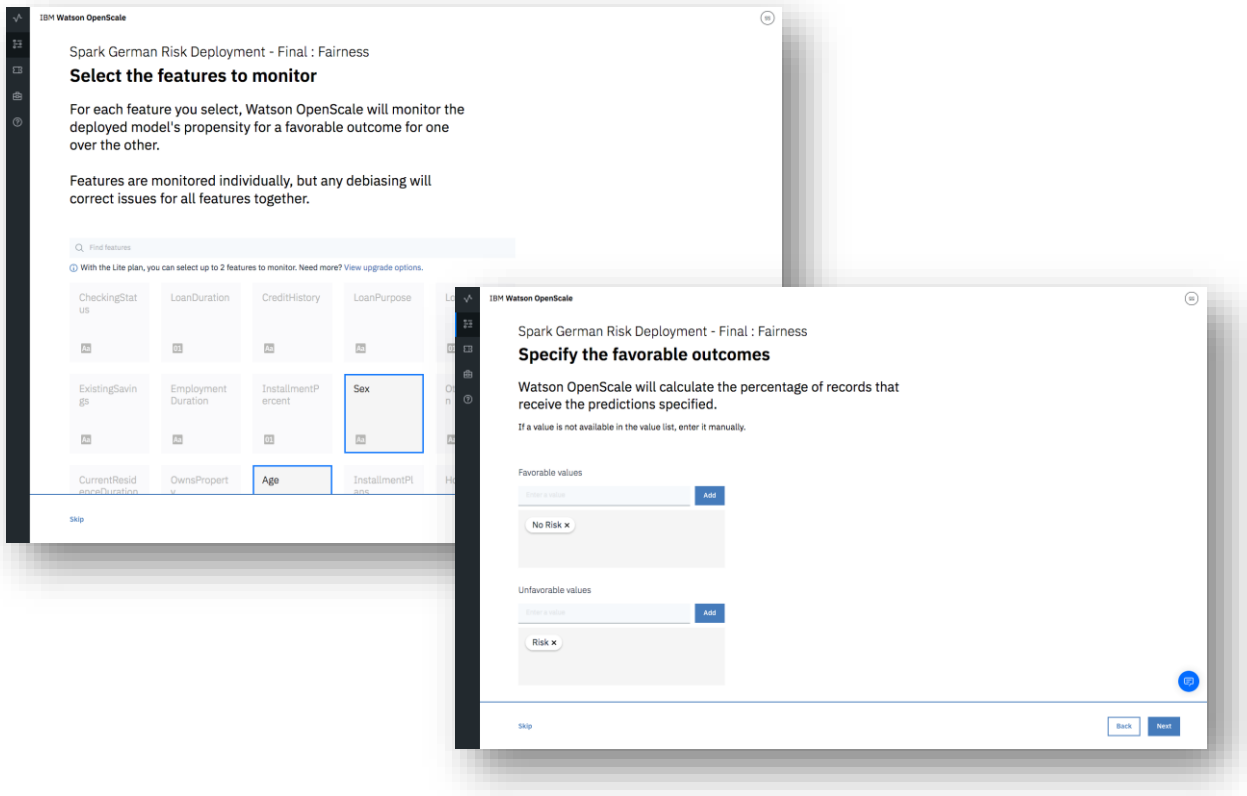
Model Fairness

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



© 2020 IBM Corporation

Explainability

Description:

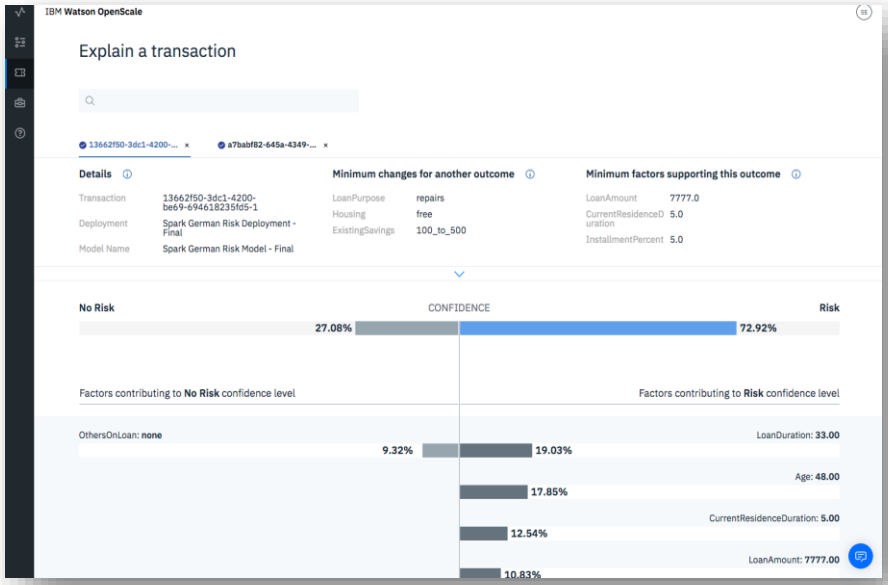
Allows you to understand which feature values of a model that are most influencing a prediction for a specific transaction

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



© 2020 IBM Corporation

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:

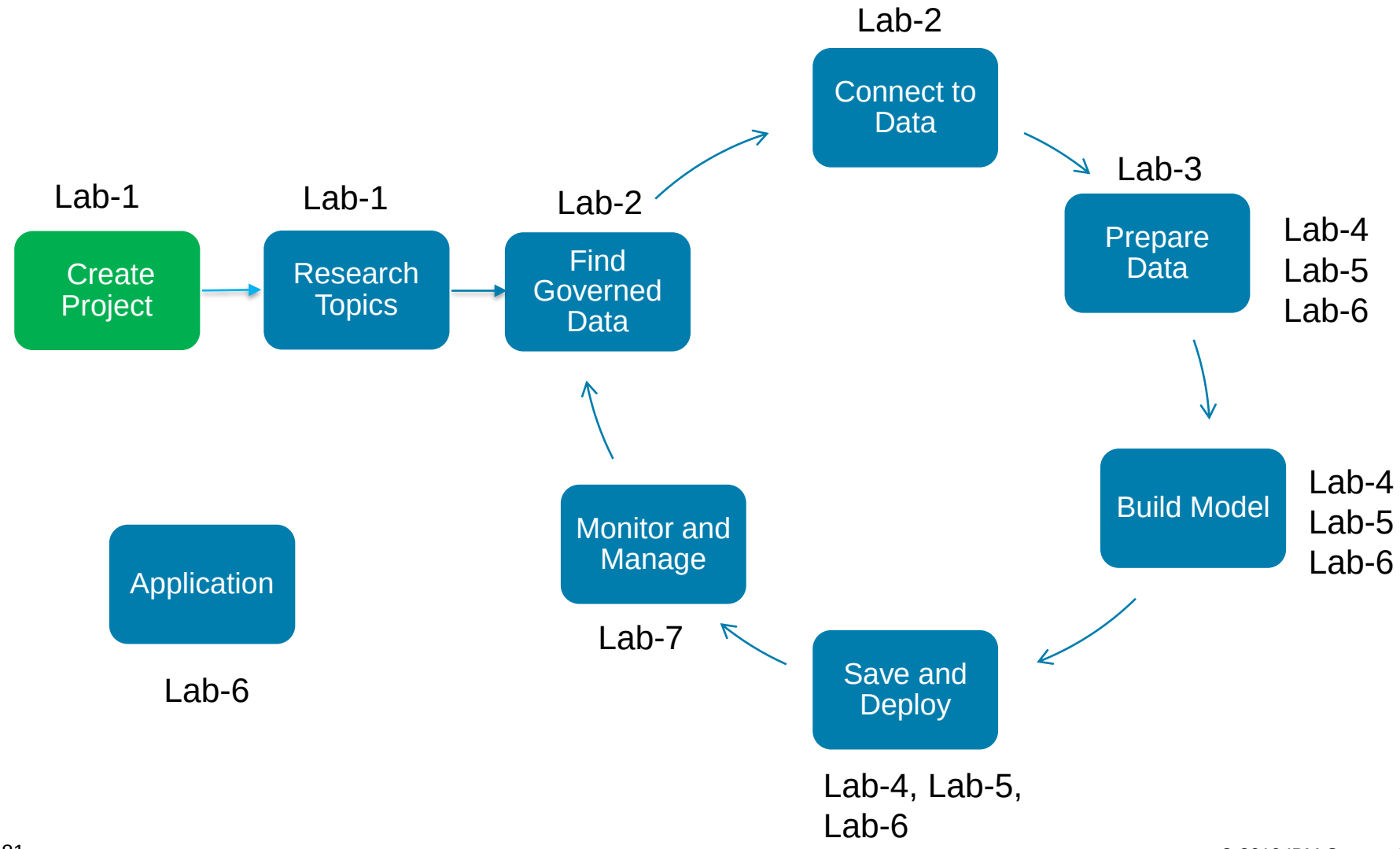
Upon completing the lab, you will

- Import and Deploy a machine learning model
- Provision a Watson OpenScale service
- Configure Watson OpenScale for Payload Logging, Quality, Fairness, and Drift.
- Submit Feedback and View Quality Metrics
- Score Data and View Fairness Metrics
- Explain a Transaction.

© 2020 IBM Corporation

Watson Studio supports the Data Science Lifecycle

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



© 2019 IBM Corporation