Hands on Introduction to IBM's Watson Studio

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

Bernie Beekman
Executive I/T Architect

Watson Studio is the new name for the IBM Data Science Experience on Cloud

Watson Knowledge Catalog is the new name for the IBM Data Catalog

Get started with Watson Studio at datascience.ibm.com

Agenda

Time	Description
1:30 PM – 2:15 PM	**Overview of Data Science and the Watson Studio Platform** **Lab Orientation**
2:15 PM – 3:00 PM	**Lab 1 - Watson Machine Learning**
3:00 PM – 3:45 PM	**Lab 2 – Watson Studio SPSS Modeler**
3:45 PM – 4:30 PM	**Lab 3 – Data Refinery**
4:30 PM – 5:15 PM	**Lab 4 - Machine Learning with SparkML and Jupyter Notebooks**
5:15 PM – 5:30 PM	**Questions and Wrap-Up**

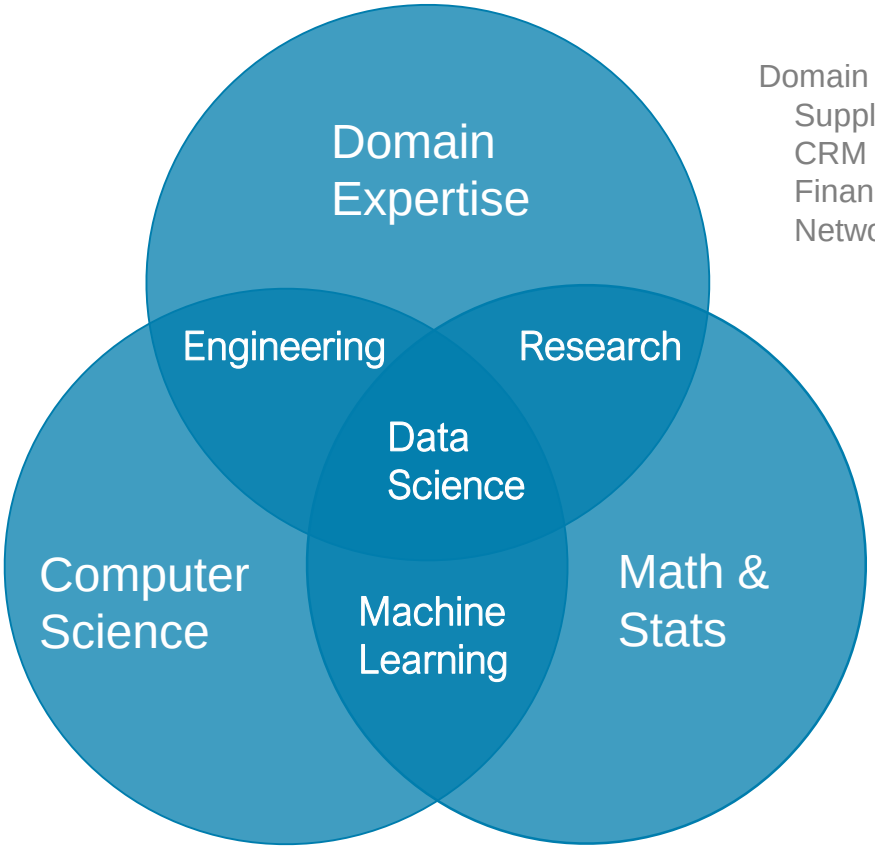
Participant Background

- R/Python/Scala
- Jupyter Notebook
- Spark
- IBM Cloud/Bluemix
- Machine Learning
- Deep Learning/Neural Networks
- Github

Outline

- Data Science Introduction
- Watson Studio Overview
- Lab Overview

What is Data Science?



Domain Knowledge
- Supply Chain
- CRM
- Financials
- Networking

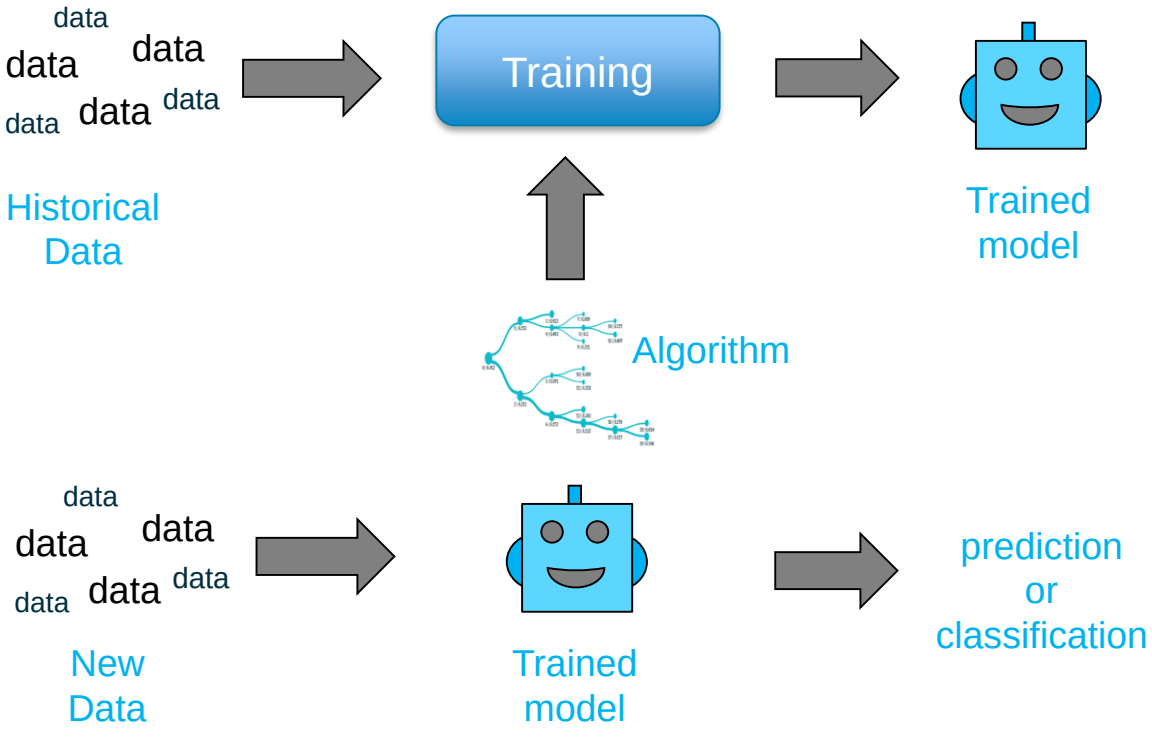
Scripting, SQL
Python, R Scala
Data Pipelines
Big Data/ Apache Spark

Mathematics
Computational

Data Science Projects Require Multiple Skills

But what is Machine Learning?

"Computers that learn without being explicitly programmed"

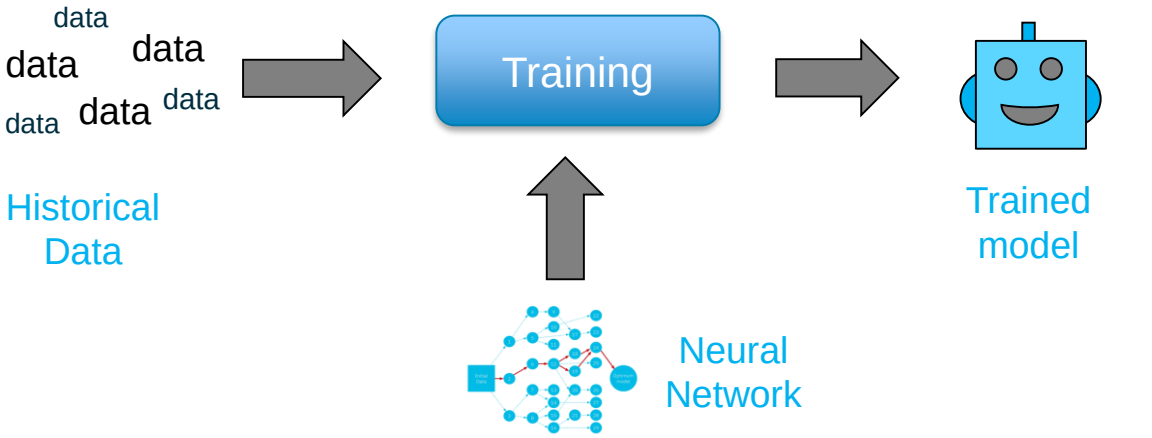


1. A machine learning model is trained to recognize patterns in historical data

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.

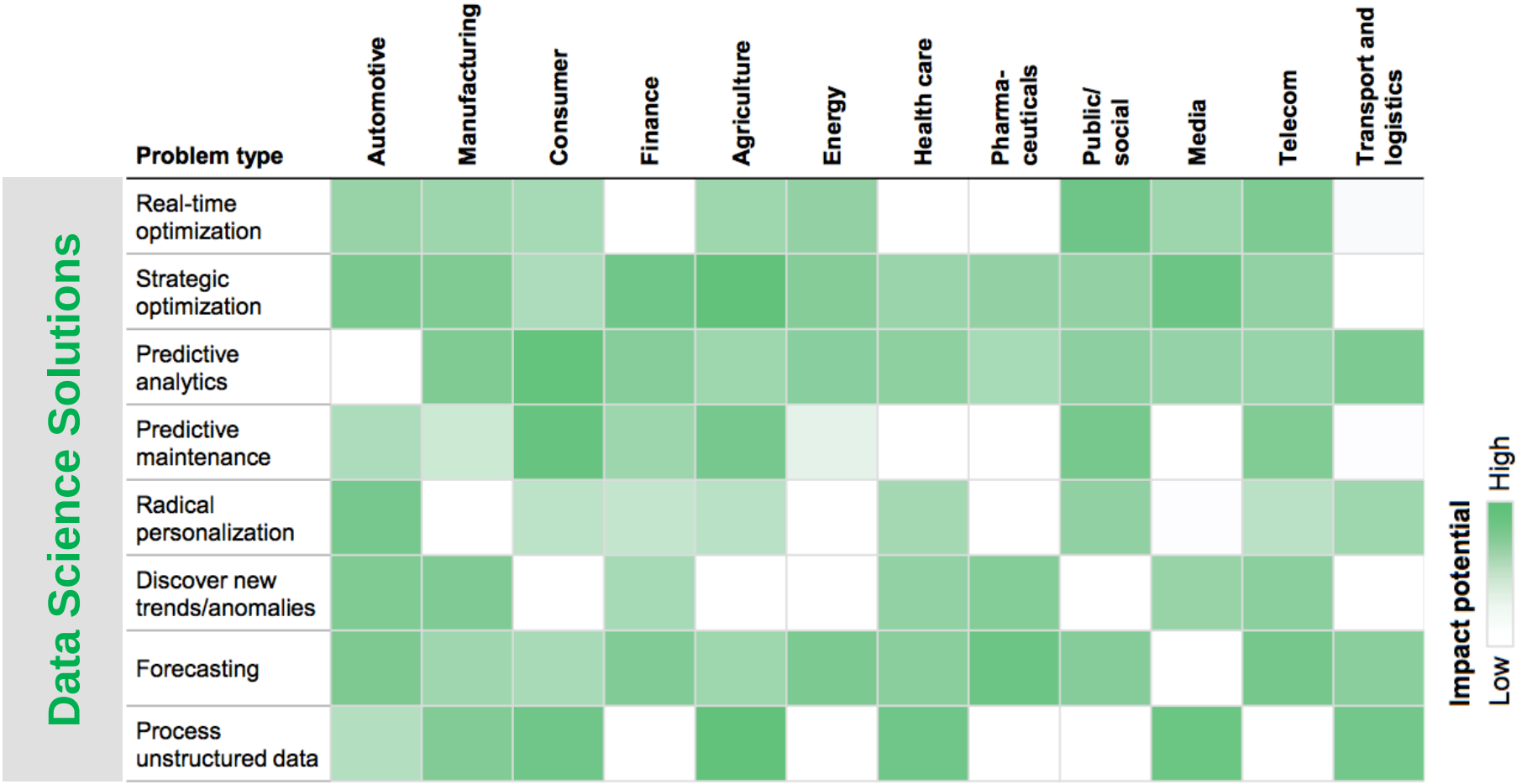
But what is Deep Learning?

"Computers that learn without being explicitly programmed"



Data Science Impact Across Industries and Use Cases

$10s of Billions in each industry and use case



SOURCE: McKinsey Global Institute analysis

Challenges in delivering value with Data Science

Data

- Data resides in silos and difficult to access
- Unstructured and external data wasn't considered

Governance

- Self-service isn't a reality, if the data isn't secure
- Understanding lineage and getting to a system of truth

Skills

- Data Science skills are in low supply and high demand

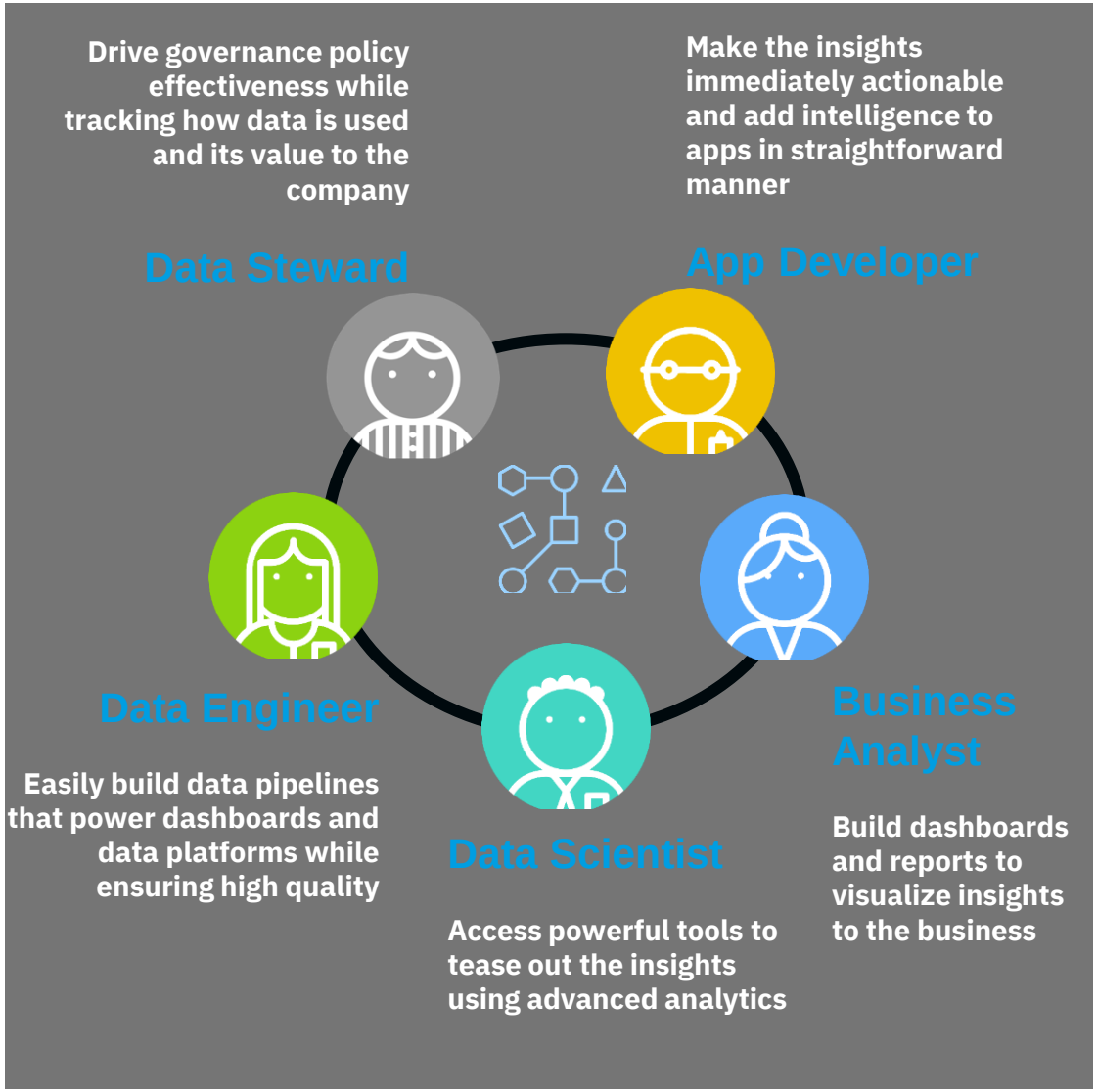
Infrastructure

- Need an environment that enables collaboration and deployment to production
- Discrete tools present barriers to progress

Watson Studio Platform

IBM Watson Studio Platform

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



Watson Studio supports end-to-end AI workflow

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

Connect and discover content from multiple data sources in the cloud or on premises. Bring **structured** and **unstructured** data to one toolkit.

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.

Clean and prepare your data with **Data Refinery**, a tool to create data preparation pipelines visually.
Use popular open source libraries to prepare unstructured data.

Democratize the creation of ML and DL models. Design your AI models **programmatically** or **visually** with the most popular **open source** and IBM ML/DL frameworks. Train at scale on **GPUs** and **distributed** compute

Deploy your models easily and have them **scale automatically** for online, batch or streaming use cases

Monitor the performance of the models in production and trigger automatic retraining and redeployment of models.

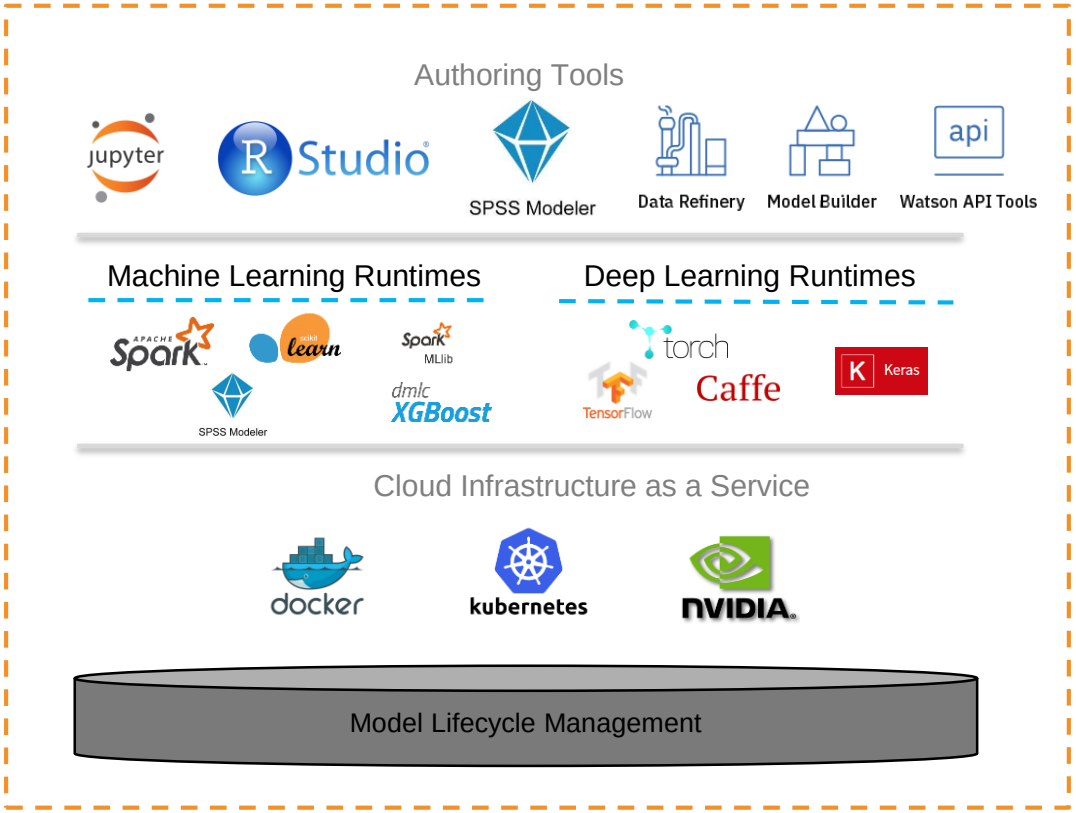
13

Watson Studio Tools

- Create, collaborate, deploy, and monitor
- Best of breed open source & IBM tools
- Code (R, Python or Scala) and no-code/visual modeling tools

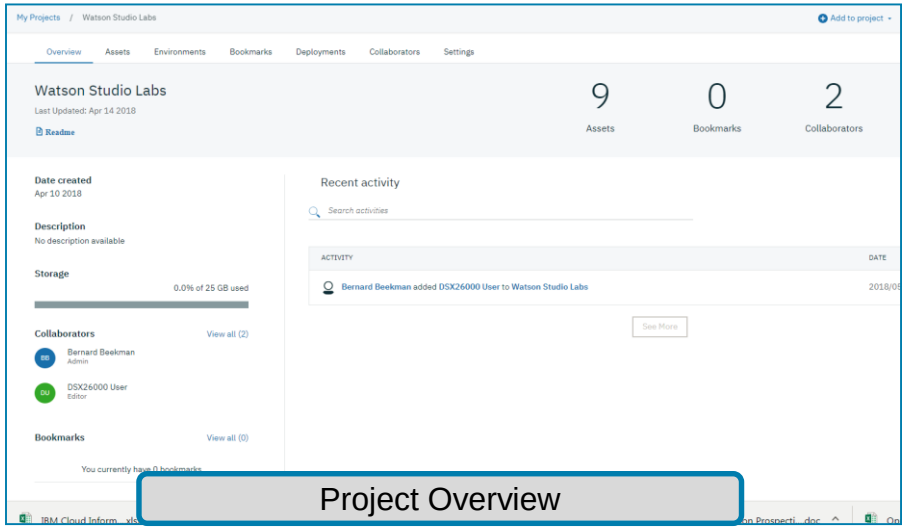
- Open Source and IBM libraries/frameworks

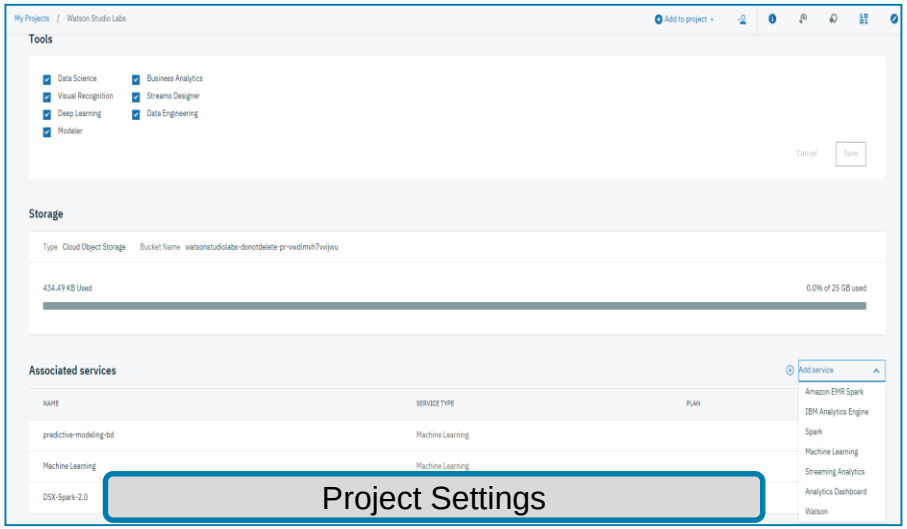
- Fully managed service
- Container-based resource management
- Elastic pay as you go cpu/gpu power

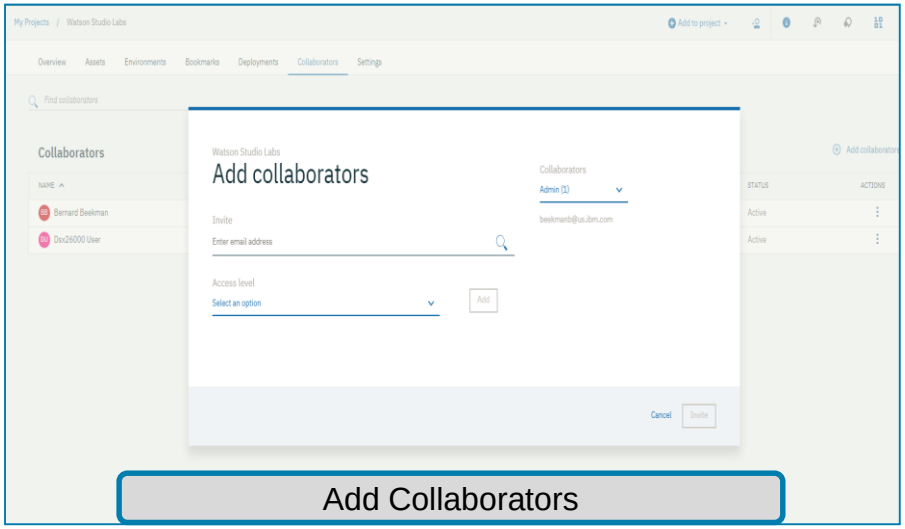


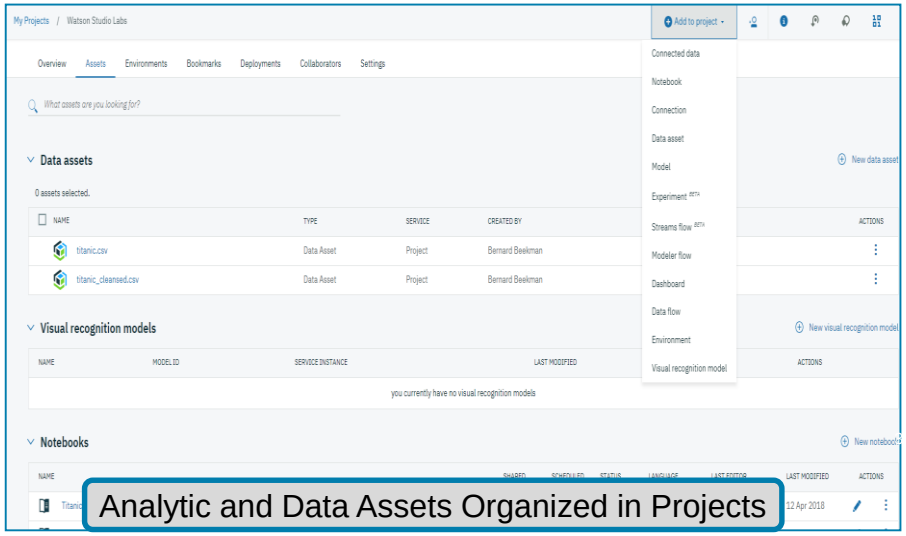
Watson Studio – Projects

Making Data Science a Team Sport



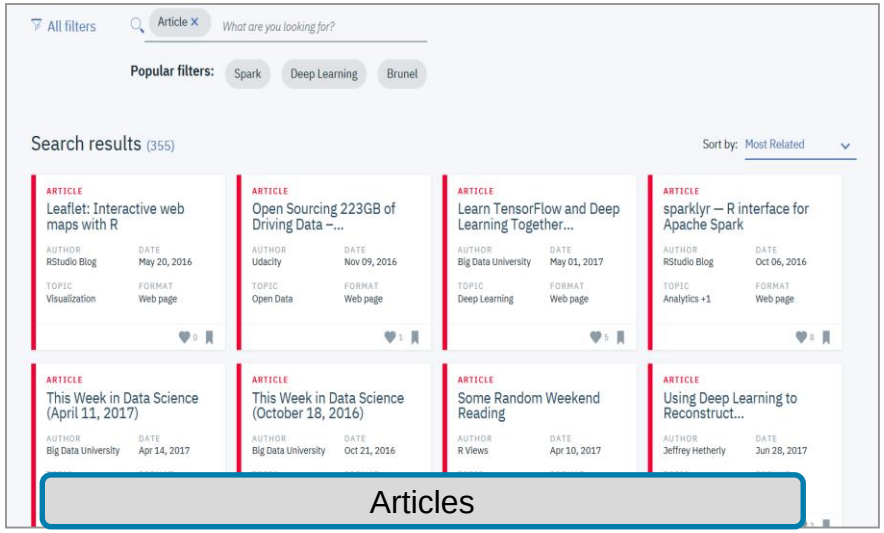


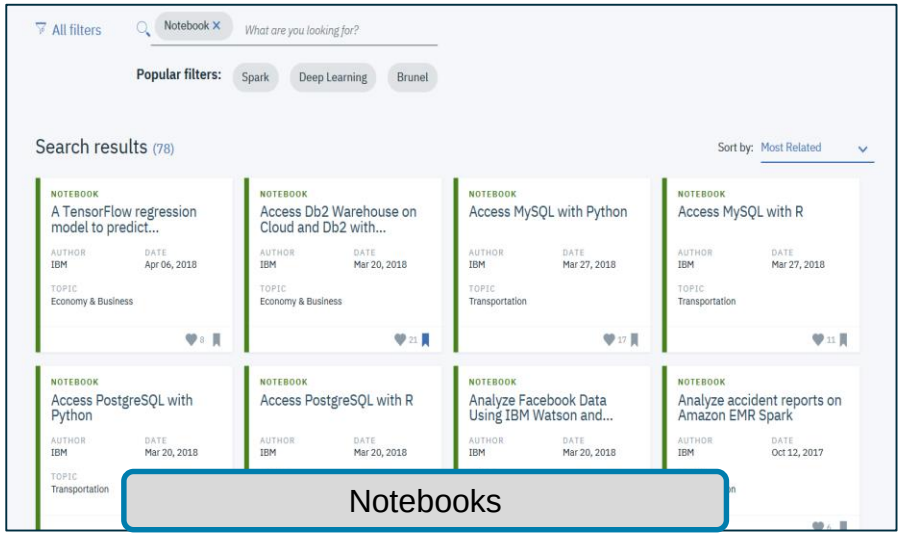


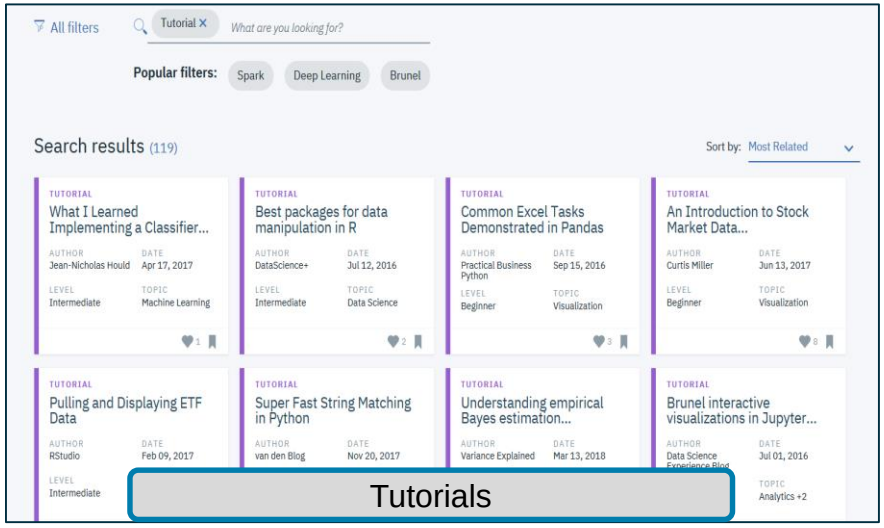


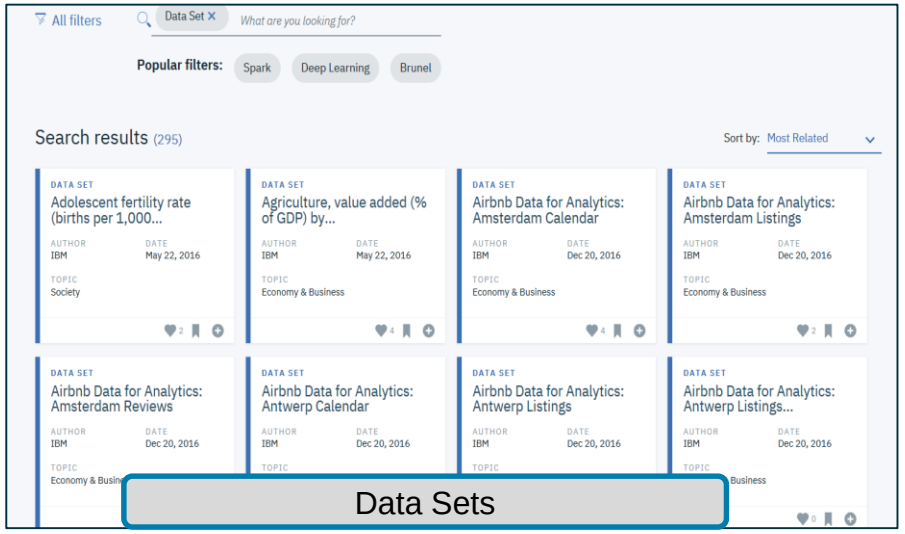
Watson Studio – Community Cards

Built-in learning to get started





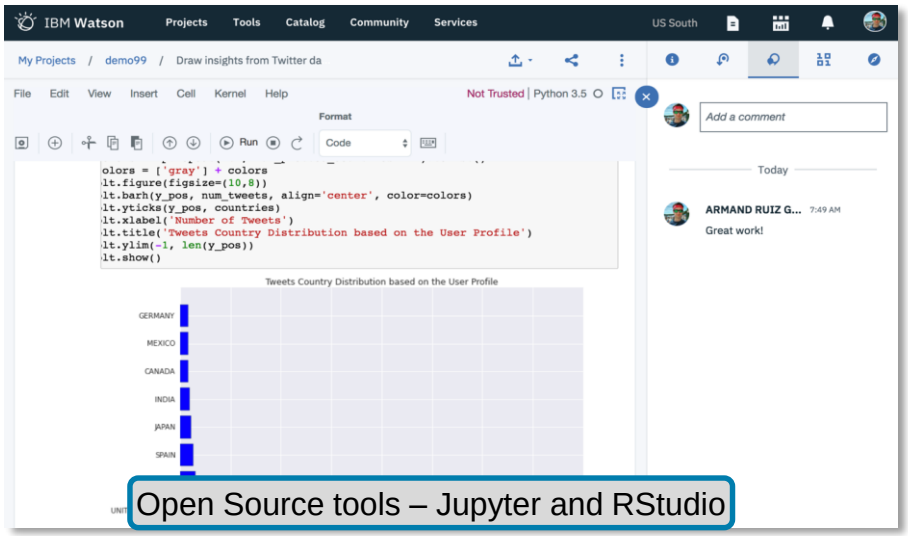




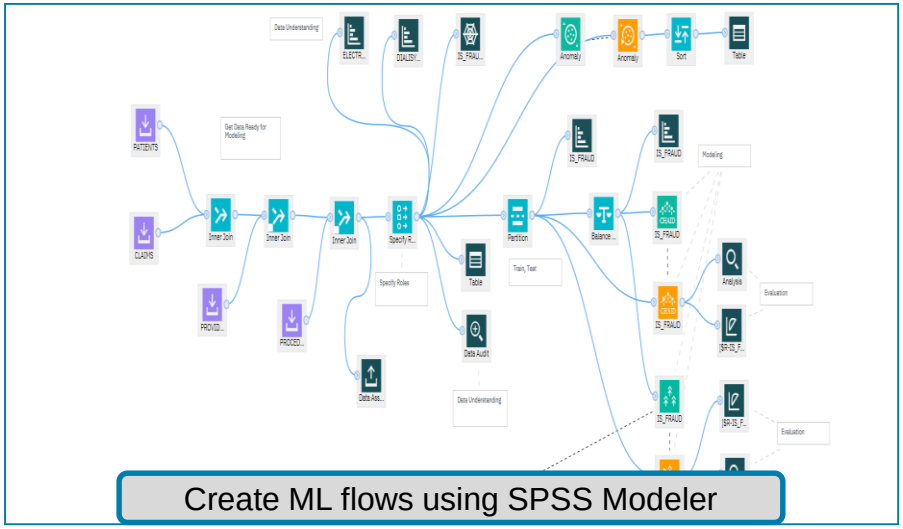
Watson Studio – Create Assets

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

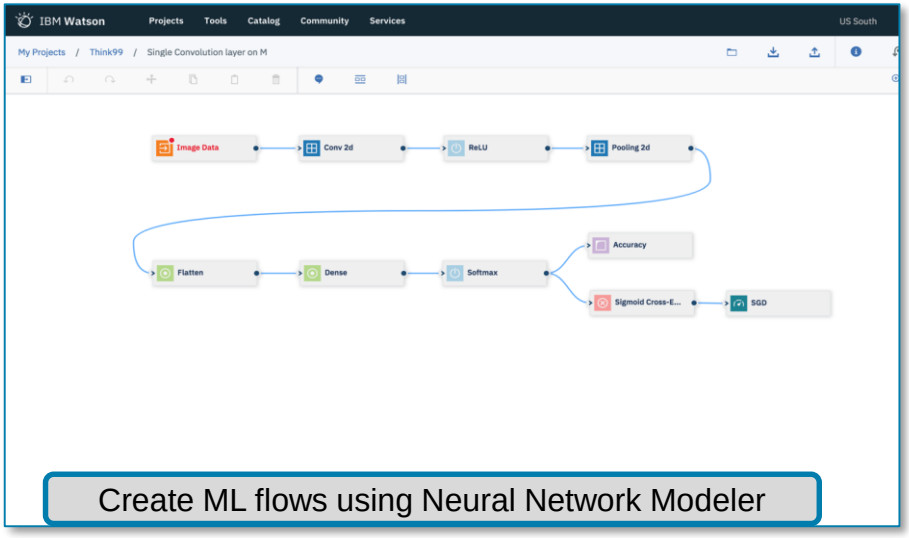
Connect to Data Sources



Open Source tools – Jupyter and RStudio



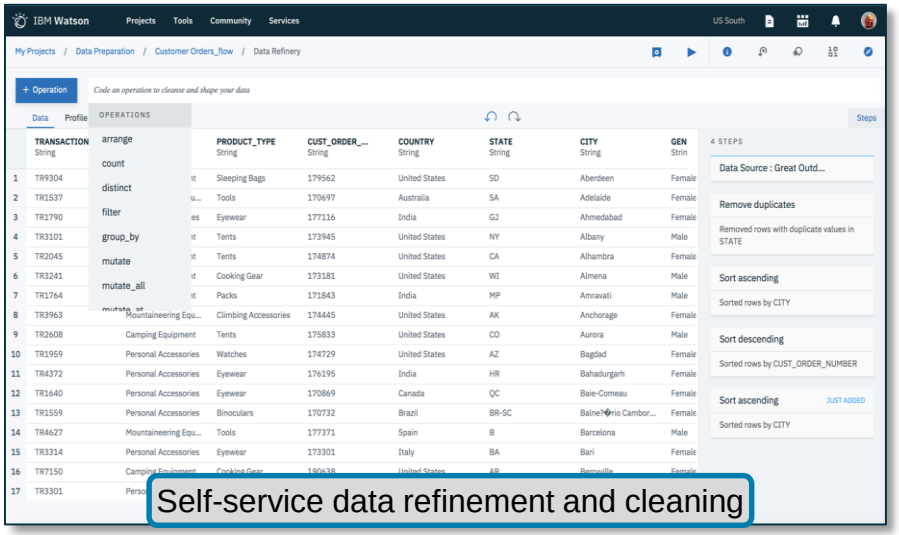
Create ML flows using SPSS Modeler



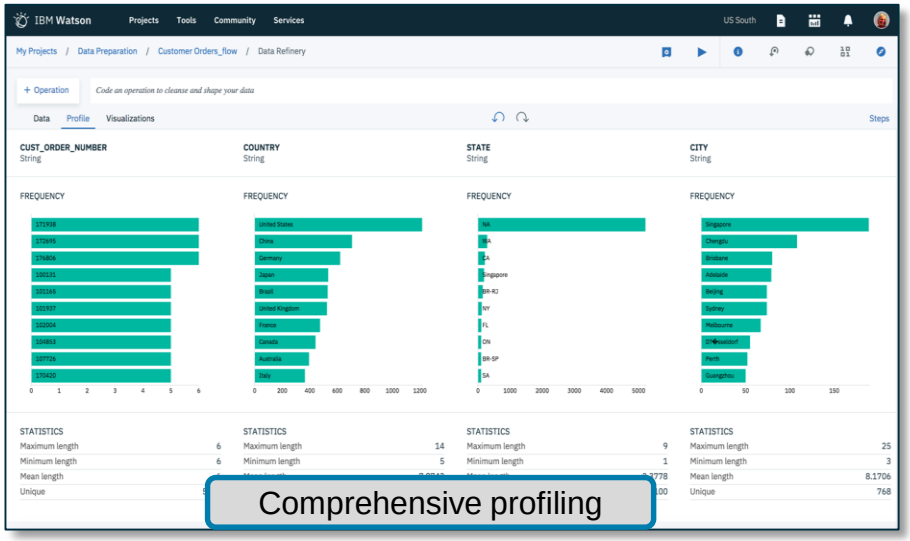
Create ML flows using Neural Network Modeler

Watson Studio - Data Refinery

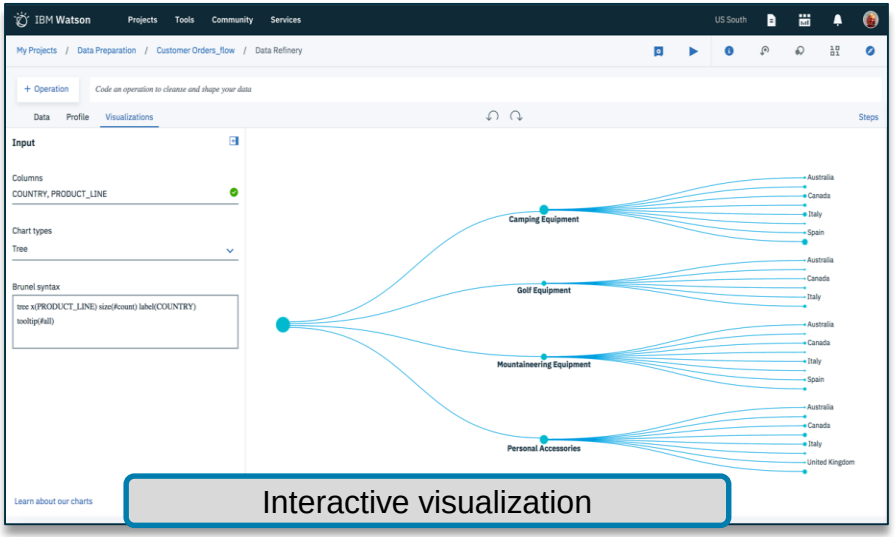
Making Data fit for use



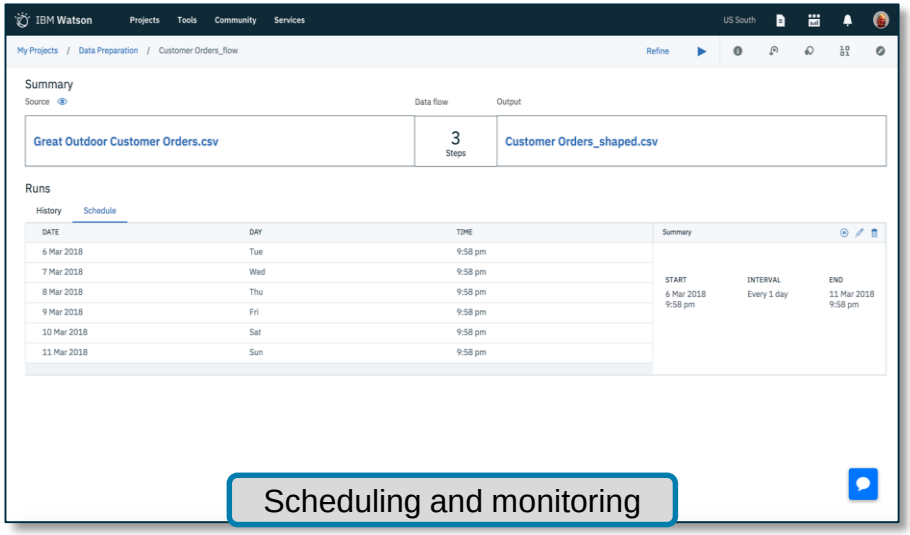
Self-service data refinement and cleaning



Comprehensive profiling



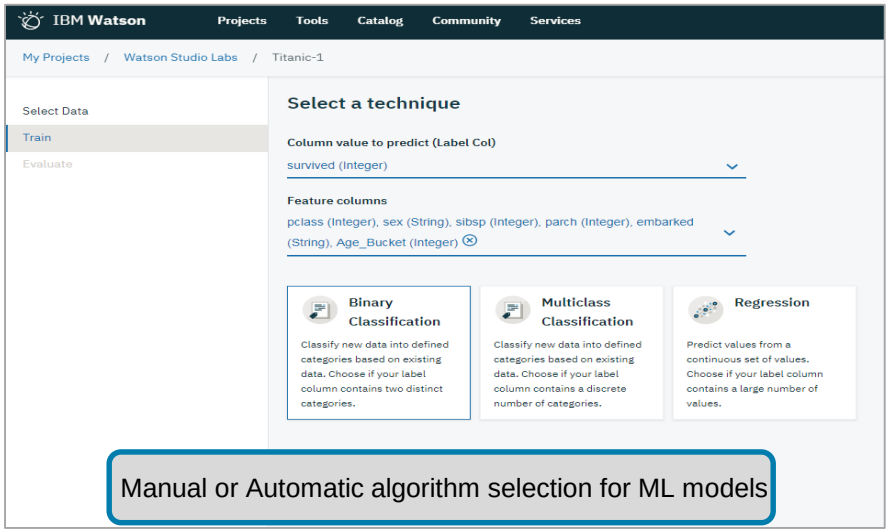
Interactive visualization



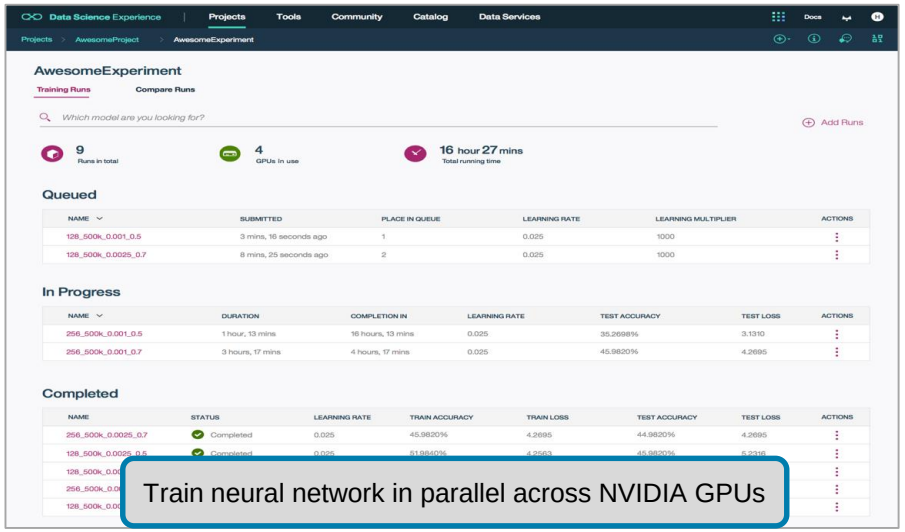
Scheduling and monitoring

Watson Studio – Watson Machine Learning

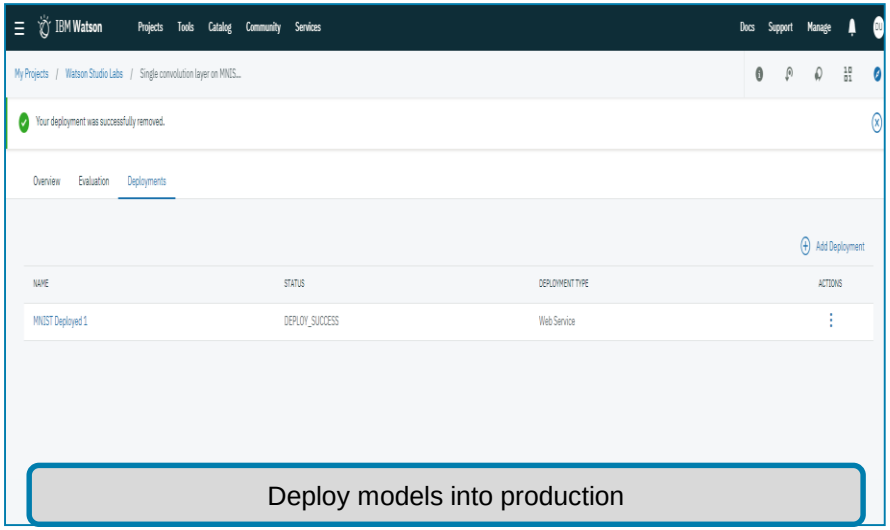
Simplifying deployment and management of ML models in production



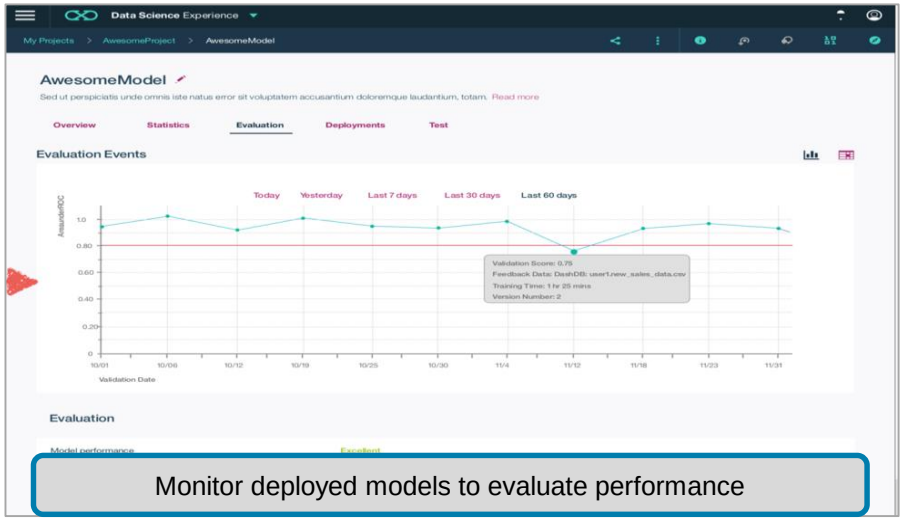
Manual or Automatic algorithm selection for ML models



Train neural network in parallel across NVIDIA GPUs



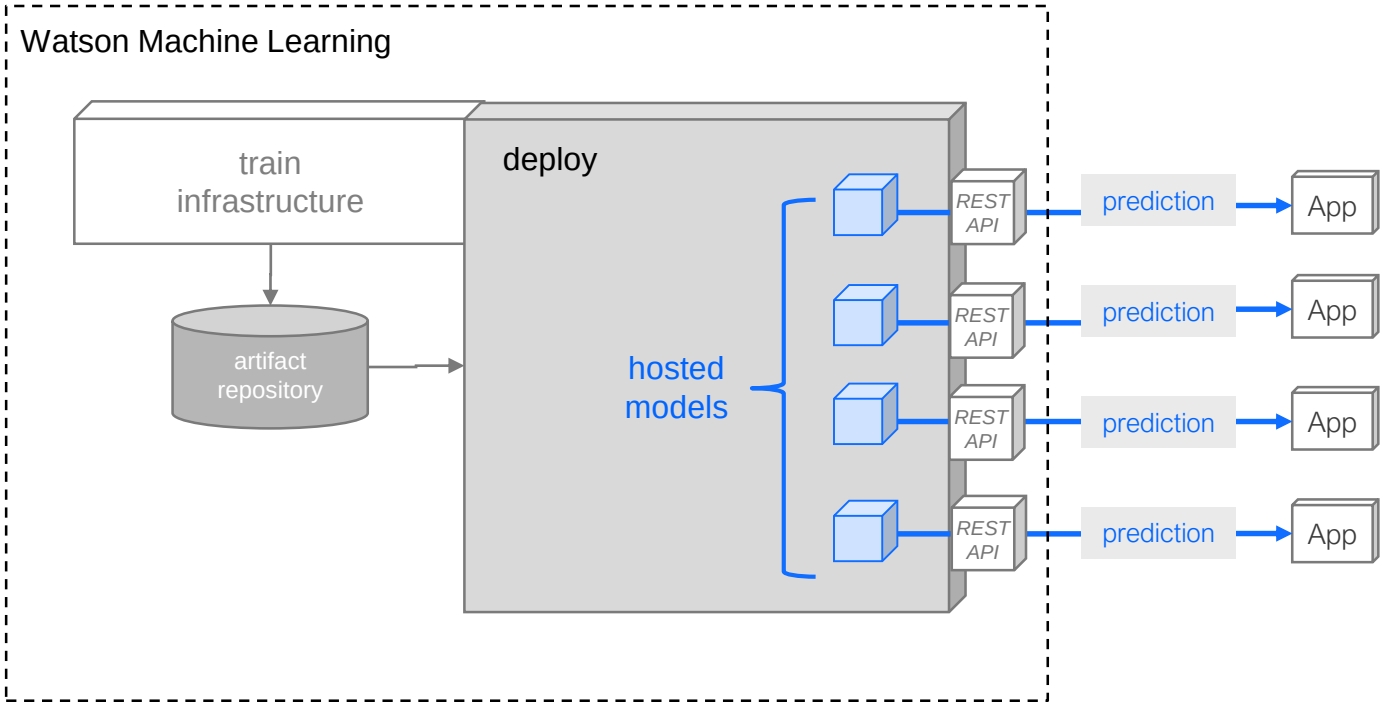
Deploy models into production



Monitor deployed models to evaluate performance

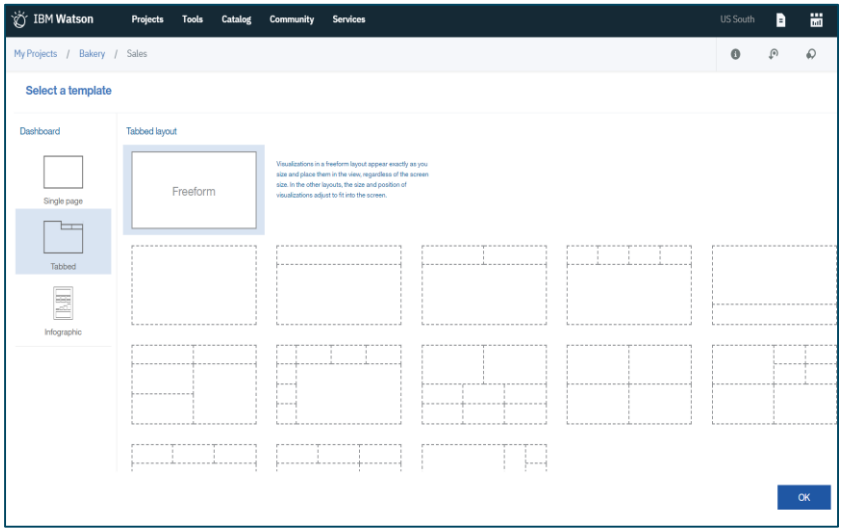
Watson Studio- Deploying Trained Models

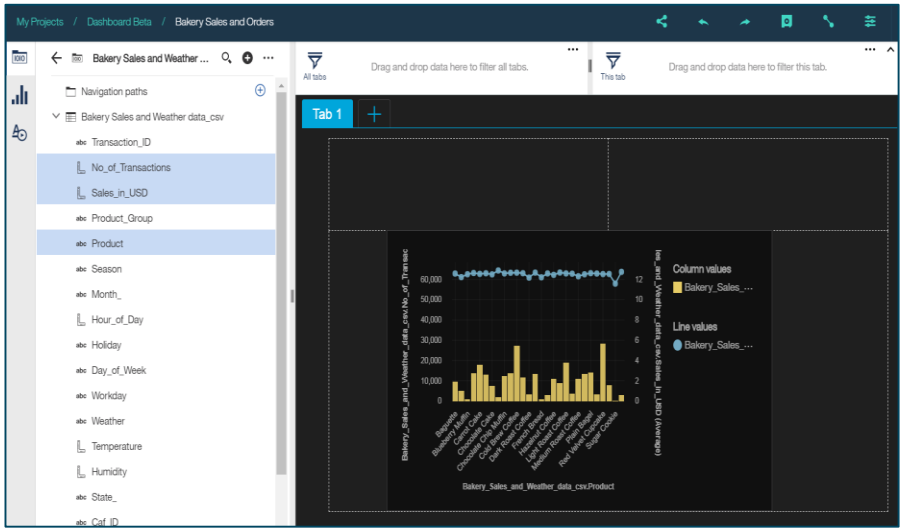
Deploy your models within Watson Machine Learning

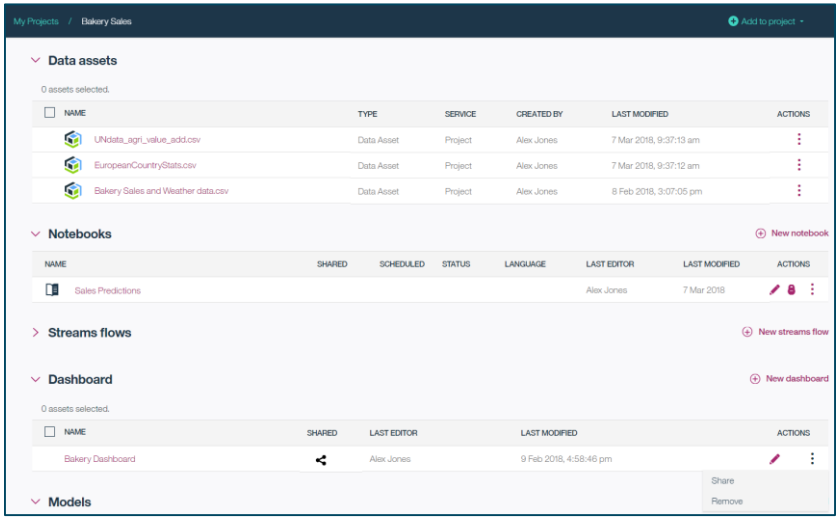


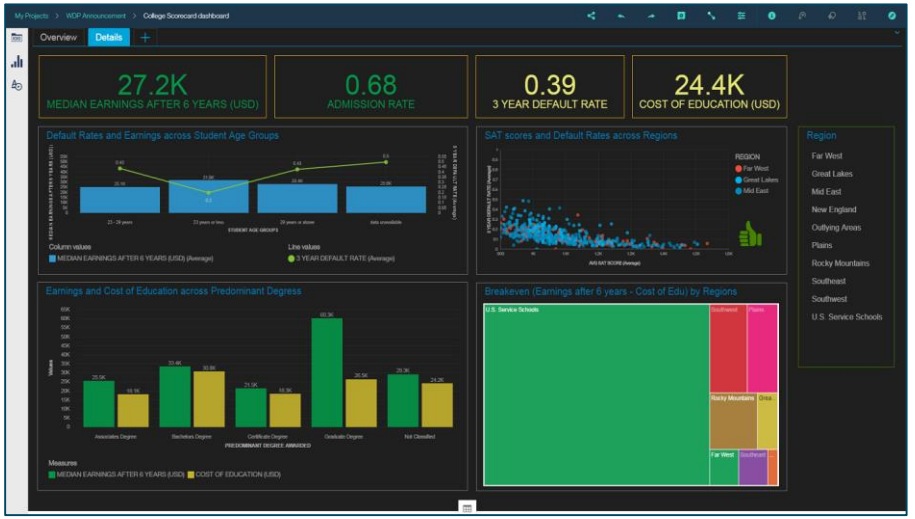
Watson Studio – Dynamic Dashboards

Making insights available to all



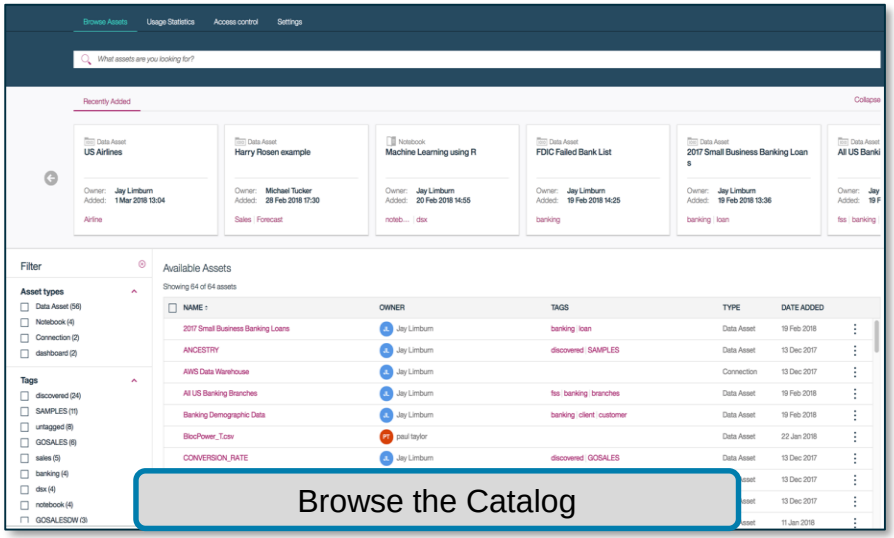




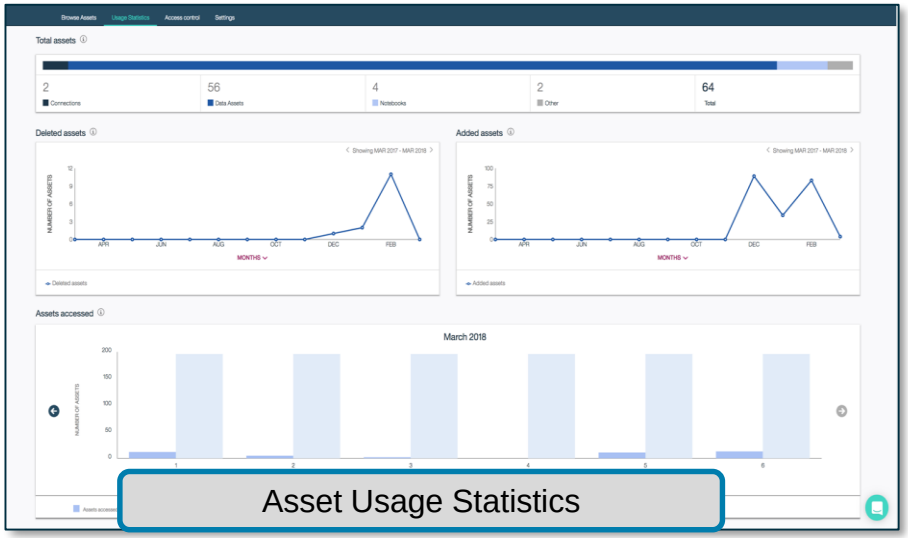


Watson Knowledge Catalog

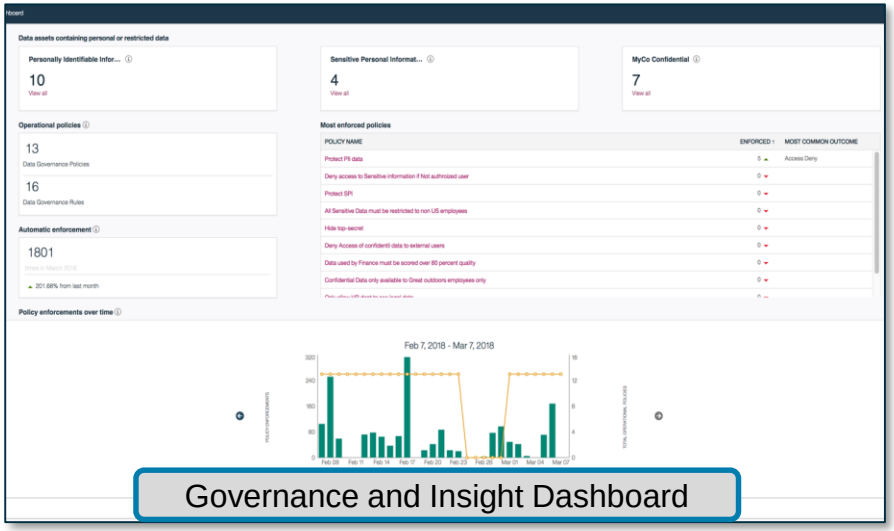
Unlock tribal knowledge and unleash knowledge workers



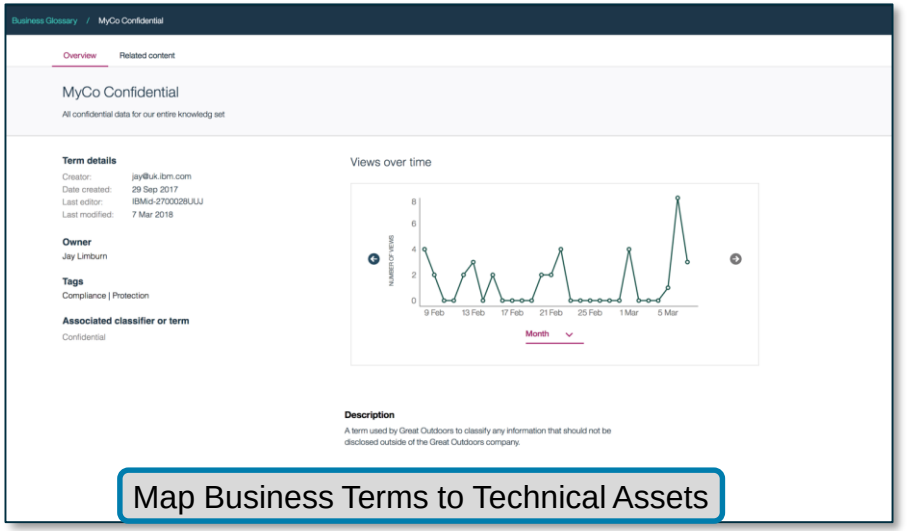
Browse the Catalog



Asset Usage Statistics



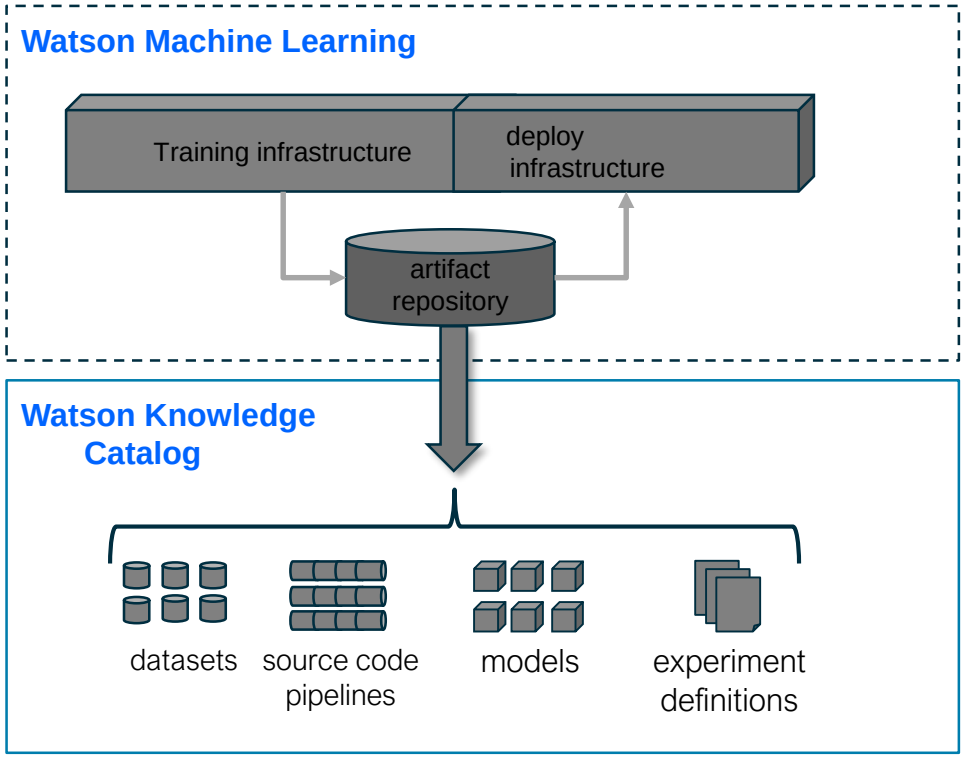
Governance and Insight Dashboard



Map Business Terms to Technical Assets

Watson Studio Model Lifecycle Management

Use the Watson Knowledge Catalog and Watson Studio to manage your AI assets or manage them yourself



Model Explanations
In May 2018, the General Data Protection Regulation (GDPR) takes effect and grants consumers the legal "right to explanation" from organizations that use algorithmic decision making.

Audit Trails
Tracking prediction to each model's unique heritage is critical to regulatory compliance. Enforcing access controls for model sharing and deployment ensure ensures data security and application stability.

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

Experiment centric DL workflow

- Monitor batch training experiments then compare cross-model performance without worrying about log transfers and scripts to visualize results.
- You focus on designing your neural networks. We'll manage and track your assets.

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

24

How does Watson Studio help fulfill the promise of your data?

Data

Puts every important data source at the fingertips of the teams that need it wherever resides

Governance

Enforces your policies without getting in the way of delivering insights

Skills

Makes the most of the data professionals you have and helps them grow and learn from each other as a team

Infrastructure

Brings all the tools in one place. Collaboration capabilities enables Data Science as a team sport.

Watson Studio is the new name for the IBM Data Science Experience on Cloud

Watson Knowledge Catalog is the new name for the IBM Data Catalog

Get started with Watson Studio at datascience.ibm.com

Lab Overview

Lab Overview

Use IBM's Watson Studio to create machine learning models and applications. Participants will be led through 4 labs (time permitting). Lab-1, Lab-2, Lab-3, and Lab-4 all use the Titanic data set, a common one used in Kaggle competitions.

- **Lab-1 - The first lab will use the Watson Machine Learning capability to create a machine learning model based on the Titanic data set. The model will be deployed in the IBM Cloud, and an application will be built that uses the deployed machine learning model to predict survivability given passenger characteristics.**
- **Lab-2 - The second lab will guide participants in using the Watson Studio SPSS Modeler capability to explore, prepare, and model passenger data from the Titanic. The SPSS Modeler is a drag and drop capability to build machine learning pipelines.**
- **Lab-3 - The third lab features the Data Refinery tool a fully managed self-service data preparation facility.**
- **Lab-4 - The fourth lab will leverage Spark machine learning (SparkML) in a Jupyter notebook to predict survivability using pyspark and a supervised learning model.**

Lab Tips

- Labs are all located in www.github.com/bleonardb3/AA repository. Environment set up is located in the repository README file. We will jointly walk through these steps.
- Instructions for each Lab are in the README file in the respective Lab folder.
- With cloud development frequent improvements are made 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.
- You need to download the pdfs that are linked to the instructions for Lab-1, Lab-2, and Lab-3. You will click on the link and then click on the Download option. Otherwise, the links in the pdf will not work when viewing in the github interface.
- When downloading csv data files, make sure you follow the instructions to right click on the Raw button and use the Save link as … option.
- Do not use Internet Explorer as the browser
- For Lab 4, you execute notebook cells by <Shift><Enter> when your cursor is in a code cell.

Lab 1 – Watson Machine Learning

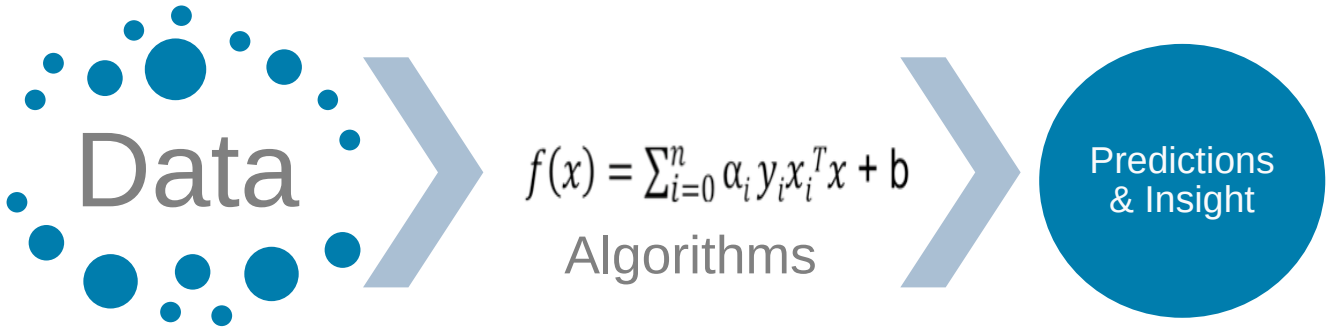
In this lab, you will use IBM's Watson Machine Learning GUI to train, evaluate, and deploy a Watson Machine Learning model based on the Titanic dataset.

Objectives:

- Upon completing the lab, you will:
  - Become familiar with the Watson Machine Learning GUI.
  - Train/Evaluate a machine learning model
  - Deploy a machine learning model.
  - Use DevOps to build and deploy an application that invokes the machine learning model service.

What is Machine Learning?

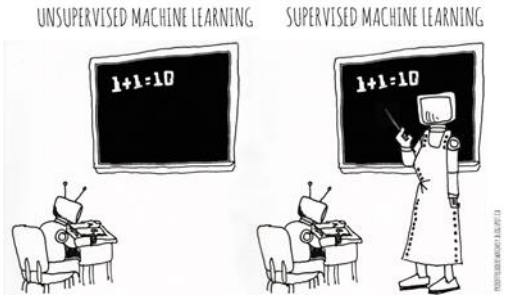
"Computers that learn without being explicitly programmed"



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

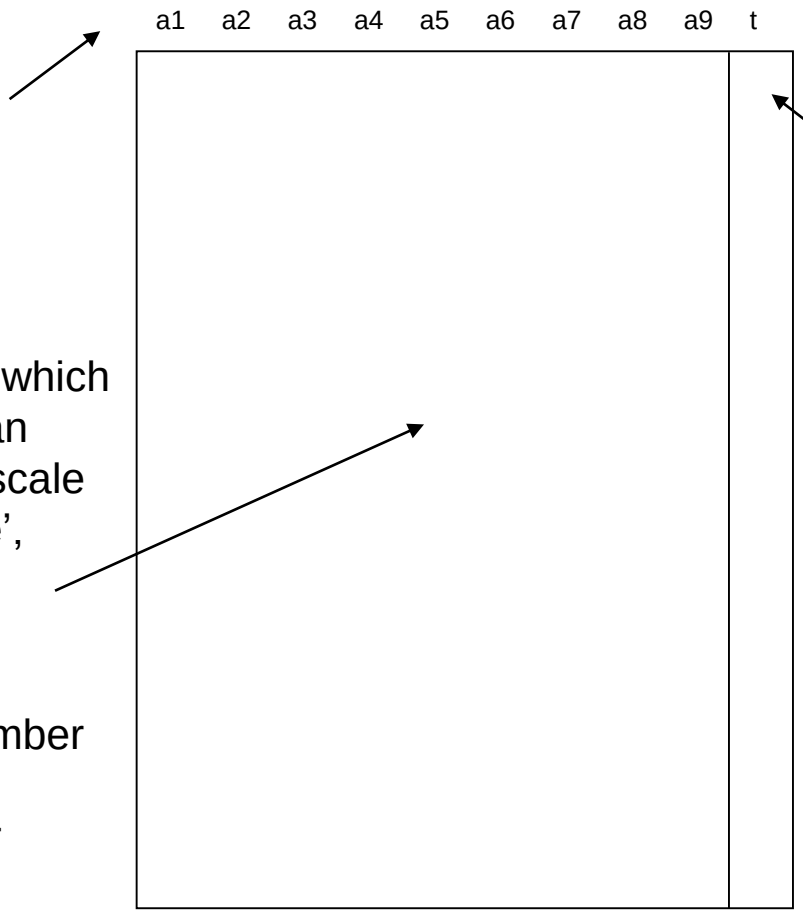


Categories of Machine Learning

Technique	Usage	Algorithms
Classification (or prediction)	• Used to predict group membership (e.g., will this employee leave?) or a number (e.g., how many widgets will I sell?)	• Decision Trees • Logistic Regression • Random Forests • **Naïve Bayes** • Linear Regression • Lasso Regression etc
Segmentation	• Used to classify data points into groups that are internally homogenous and externally heterogeneous. • Identify cases that are unusual	• K-means • Gaussian Mixture • Latent Dirichlet allocation etc
Association	• Used to find events that occur together or in a sequence (e.g., market basket)	• FP Growth

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

Demo Data - Titanic

- **Variable Descriptions:**

```
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

Lab 2 – SPSS Modeler

In this lab, you will use the Watson Studio SPSS Modeler capability to explore, prepare, and model passenger data from the Titanic. 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 Titanic data set
  - Explore the Titanic data set with visualizations
  - Cleanse and Transform the data
  - Train/Evaluate a machine learning mode.

Lab 3 – Data Refinery

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 to help determine missing values
  - 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 – Jupyter Notebook and SparkML

In this lab, you will use IBM's Watson Studio to create a Jupyter notebook to examine the principles of Spark Machine Learning using the Titanic dataset. You will build a model to predict who survived -- and who did not.

Objectives:

- Upon completing the lab, you will know how to:
  - Create a Jupyter notebook from a URL
  - Load data from a URL
  - Examine the data in PixieDust
  - Examine and shape the data for use in an ML model
  - Build a Pipeline for a Logistic Regression model
  - Tune the model for maximal effectiveness

Spark ML

- **Spark ML is Spark's machine learning (ML) library**
- **Goal is to make machine learning scalable and easy**
  - No need to understand the detailed math!
- **Divides into two packages:**
  - spark.mllib contains the original API built on top of RDDs
  - spark.ml provides higher-level API built on top of DataFrames for constructing ML pipelines
  - A pipeline is a series of stages where each stage either transforms, or runs through a machine learning algorithm.
- **Using spark.ml is recommended because with DataFrames the API is more versatile and flexible**
  - spark.mllib will continue to be supported

Spark ML Pipeline Terminology

Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow

- **DataFrame**: Spark ML uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types
- **Transformer**: A Transformer is an algorithm which can transform one DataFrame into another DataFrame
- **Estimator**: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer
- **Pipeline**: A Pipeline chains multiple Transformers and Estimators together in a sequence to specify an ML workflow
- **Parameter**: All Transformers and Estimators share a common API for specifying parameters

Demo Flow

- **Read in Titanic dataset as a DataFrame**
  - Drop unwanted columns and rows with null or invalid data
  - Label the data (“Survived”)
- **Data Analysis**
  - Visualizations
- **Feature Engineering**
  - StringIndexer (Sex and Embarked variables)
  - Bucketizer (Age and Fare variables)
  - VectorAssembler
  - Normalizer
- **Create a Pipeline**
- **Split Ratings data into Training (80%) and Test (20%) datasets**
  - Cache the resulting DataFrames

Demo Flow (continued)

- **Fit the Pipeline to the Test data set**
  - Logistic Regression
- **Evaluate the resulting predictions**
  - Area under the ROC curve
- **Tune the model (hyperparameters)**
  - Build Parameter Grid
  - Cross-evaluate to find the best model
- **Make improved predictions using the cross-validated model**
- **Make prediction on an imaginary passenger**
- **Show how to easily reuse completed work using a different machine learning algorithm**
  - Random Forest