

# Hands on Introduction to IBM's Watson Studio



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

**Bernie Beekman**  
**Michael Cronk**  
**Aaron McKay**

**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](https://datascience.ibm.com)

# Agenda

Time	Description
8:30 AM - 9:00 AM	Registration and Coffee
9:00 AM - 10:00 AM	Overview of Watson Studio Platform Lab Orientation
10:00 AM - 11:30 AM	Lab 1 - Machine Learning with Spark ML
11:30 AM - 12:30 PM	Lab 2 – R, Shiny, and GUI Interfaces
12:30 PM - 1:30 PM	Lunch Provided
1:30 PM - 2:30 PM	Lab 3 - Build, Train, Deploy a Neural Network Model
2:30 PM - 4:00 PM	Lab 4 - Choose From Three Options
4:00 PM - 4:30 PM	Questions and Wrap Up

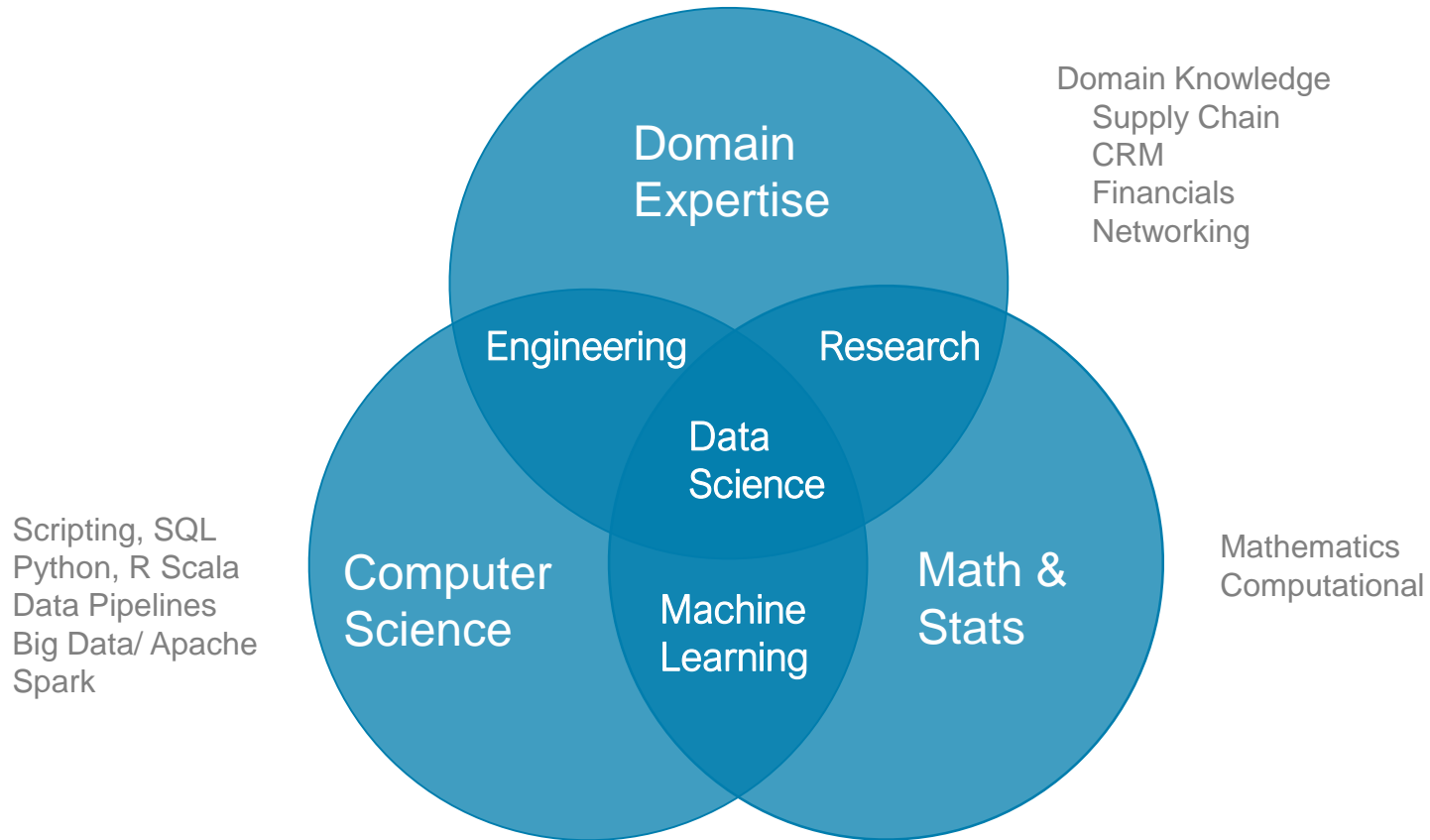
## Participant Background

- 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**

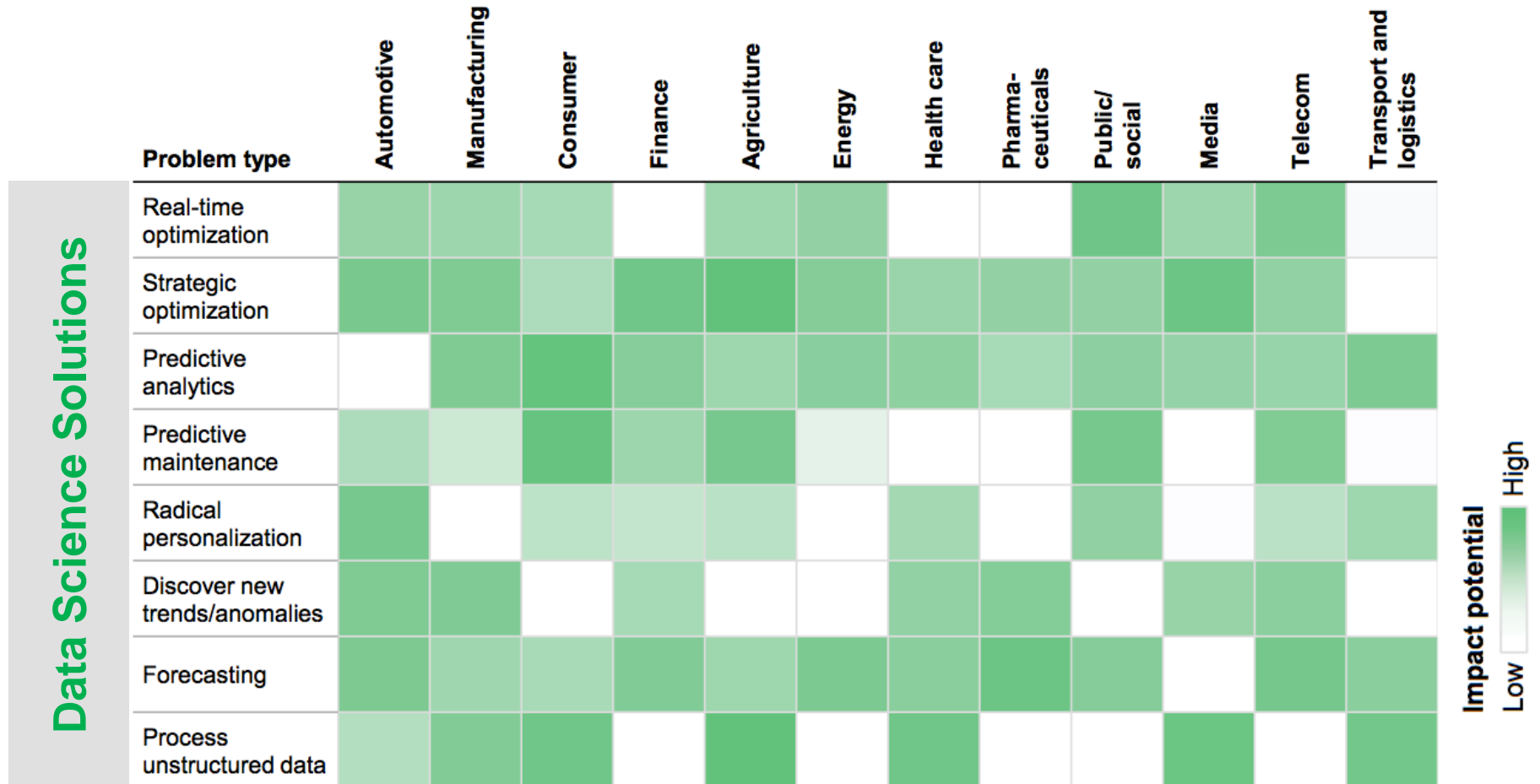
# What is Data Science?



*Data Science Projects Require Multiple Skills*

# 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

## Skills

- Data Science skills are in low supply and high demand

## Governance

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

## Infrastructure

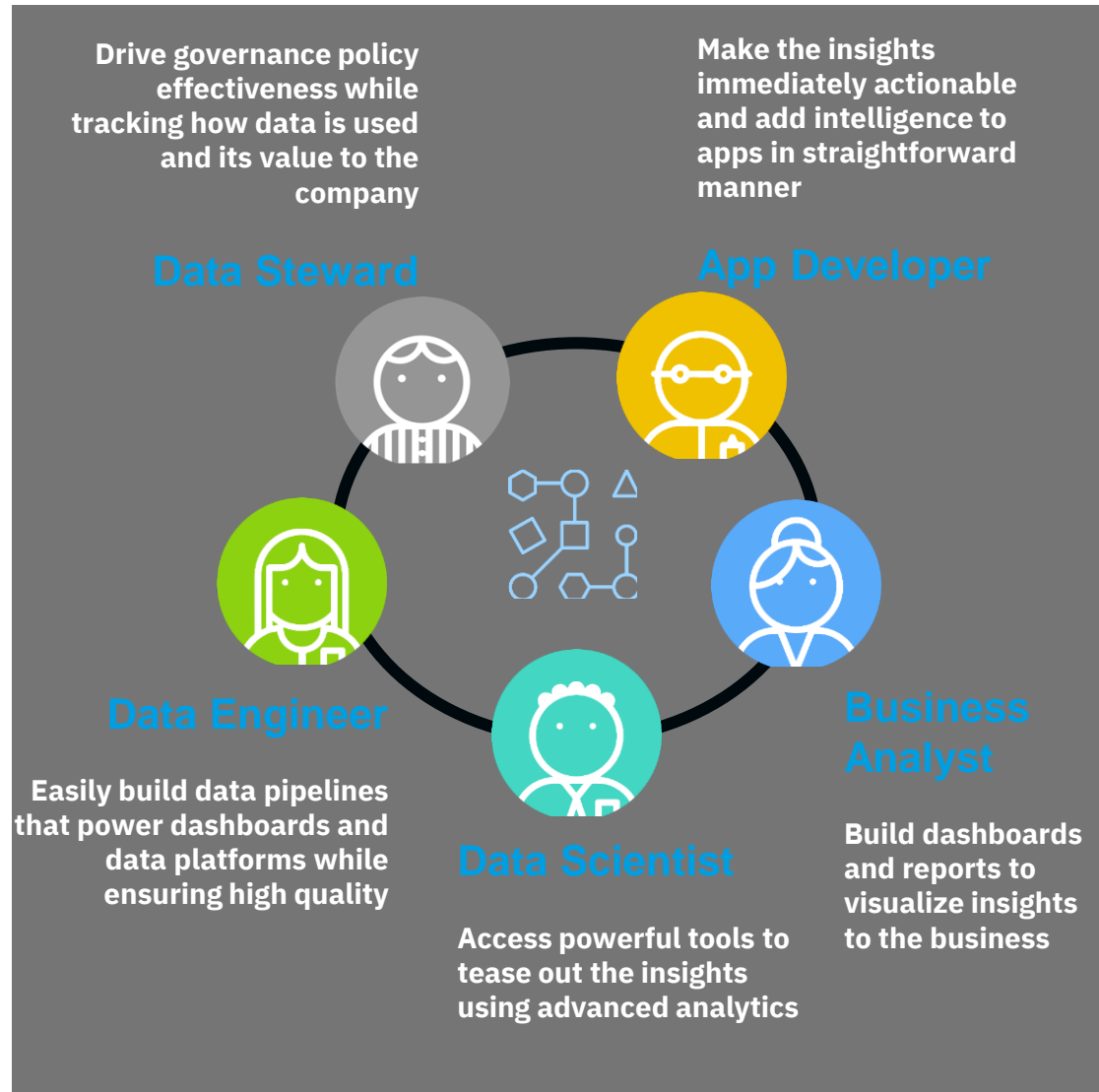
- Need an environment that enables collaboration and deployment to production
- Discrete tools present barriers to progress



# 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 &  
Access Data

Search and Find  
Relevant Data

Prepare Data  
for Analysis

Build and Train  
ML/DL Models

Deploy Models

Monitor, Analyze  
and Manage

**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.

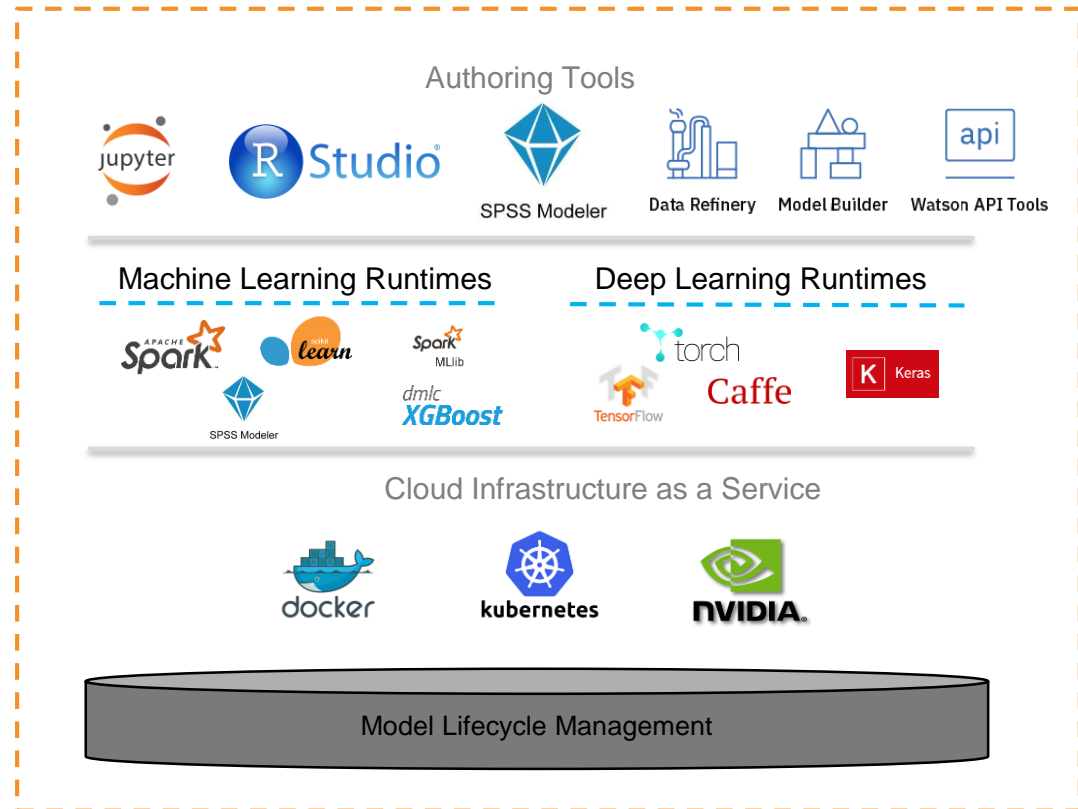
**Democratize** the creation of ML and DL models. Design your AI models **programmatically** or **visually** with the most popular **open source** and IBM ML/DL frameworks. Train at scale on **GPUs** and **distributed** compute

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.

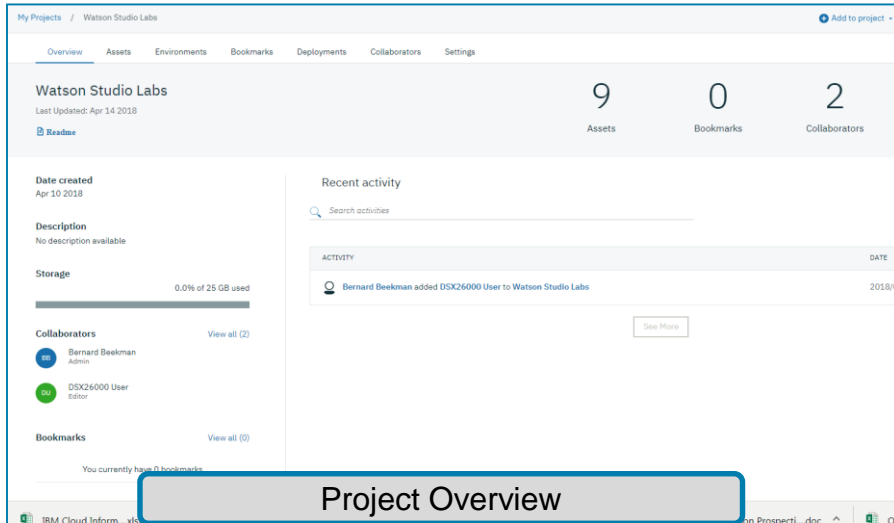
# Watson Studio Tools

- 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

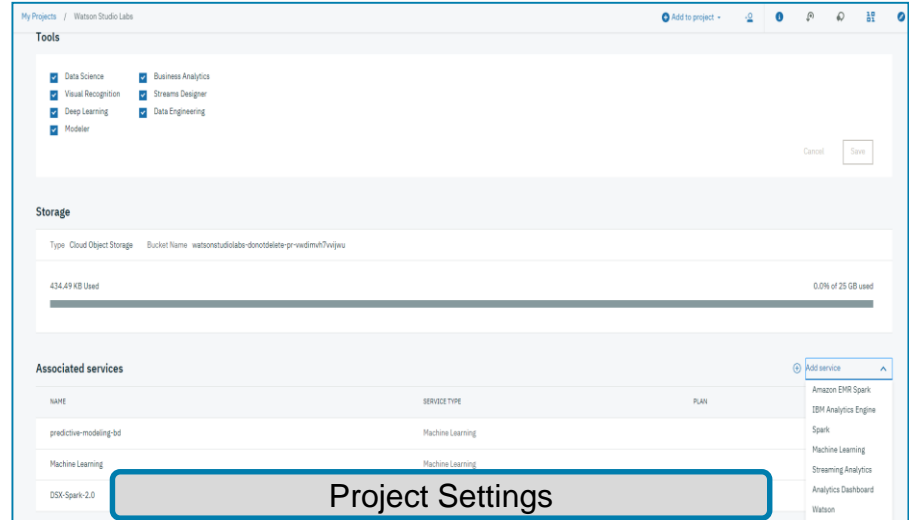


# Watson Studio – Projects

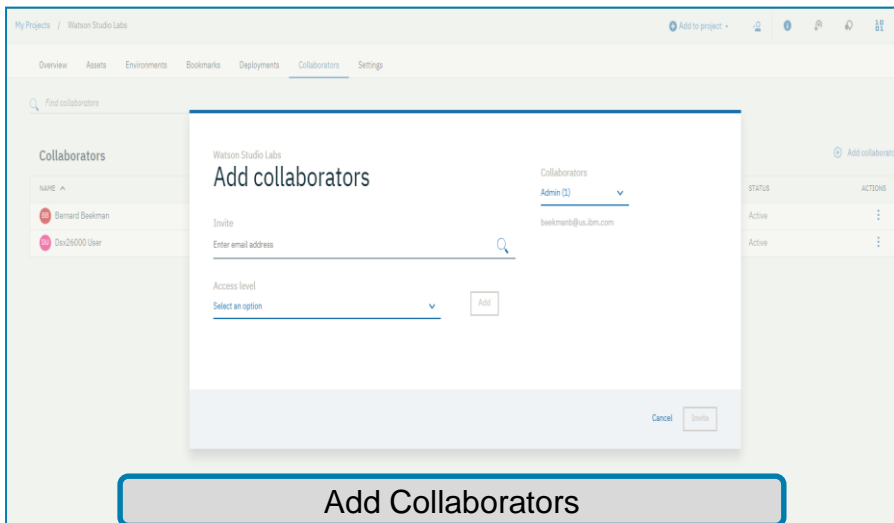
## Making Data Science a Team Sport



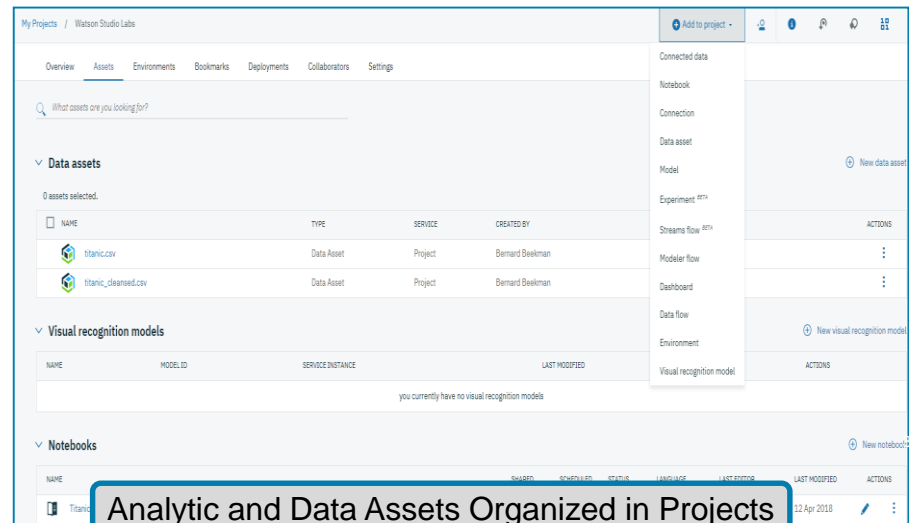
**Project Overview**



**Project Settings**



**Add Collaborators**



**Analytic and Data Assets Organized in Projects**

# Watson Studio – Community Cards

*Built-in learning to get started*

Search results (355) Sort by: Most Related

Popular filters: Spark Deep Learning Brunel

ARTICLE	ARTICLE	ARTICLE	ARTICLE
Leaflet: Interactive web maps with R	Open Sourcing 223GB of Driving Data –...	Learn TensorFlow and Deep Learning Together...	sparklyr – R interface for Apache Spark
AUTHOR: RStudio Blog	AUTHOR: Udashly	AUTHOR: Big Data University	AUTHOR: RStudio Blog
DATE: May 20, 2016	DATE: Nov 09, 2016	DATE: May 01, 2017	DATE: Oct 06, 2016
TOPIC: Visualization	TOPIC: Open Data	TOPIC: Deep Learning	TOPIC: Analytics +1
FORMAT: Web page	FORMAT: Web page	FORMAT: Web page	FORMAT: Web page
ARTICLE	ARTICLE	ARTICLE	ARTICLE
This Week in Data Science (April 11, 2017)	This Week in Data Science (October 18, 2016)	Some Random Weekend Reading	Using Deep Learning to Reconstruct...
AUTHOR: Big Data University	AUTHOR: Big Data University	AUTHOR: R Views	AUTHOR: Jeffrey Hetherly
DATE: Apr 14, 2017	DATE: Oct 21, 2016	DATE: Apr 10, 2017	DATE: Jun 28, 2017

Articles

Search results (78) Sort by: Most Related

Popular filters: Spark Deep Learning Brunel

NOTEBOOK	NOTEBOOK	NOTEBOOK	NOTEBOOK
A TensorFlow regression model to predict...	Access Db2 Warehouse on Cloud and Db2 with...	Access MySQL with Python	Access MySQL with R
AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM
DATE: Apr 06, 2018	DATE: Mar 20, 2018	DATE: Mar 27, 2018	DATE: Mar 27, 2018
TOPIC: Economy & Business	TOPIC: Economy & Business	TOPIC: Transportation	TOPIC: Transportation
NOTEBOOK	NOTEBOOK	NOTEBOOK	NOTEBOOK
Access PostgreSQL with Python	Access PostgreSQL with R	Analyze Facebook Data Using IBM Watson and...	Analyze accident reports on Amazon EMR Spark
AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM
DATE: Mar 20, 2018	DATE: Mar 20, 2018	DATE: Mar 20, 2018	DATE: Oct 12, 2017
TOPIC: Transportation	TOPIC: Transportation	TOPIC: Transportation	TOPIC: Transportation

Notebooks

Search results (119) Sort by: Most Related

Popular filters: Spark Deep Learning Brunel

TUTORIAL	TUTORIAL	TUTORIAL	TUTORIAL
What I Learned Implementing a Classifier...	Best packages for data manipulation in R	Common Excel Tasks Demonstrated in Pandas	An Introduction to Stock Market Data...
AUTHOR: Jean-Nicholas Houde	AUTHOR: DataScience+	AUTHOR: Practical Business Python	AUTHOR: Curtis Miller
DATE: Apr 17, 2017	DATE: Jul 12, 2016	DATE: Sep 15, 2016	DATE: Jun 13, 2017
LEVEL: Intermediate	LEVEL: Intermediate	LEVEL: Beginner	LEVEL: Beginner
TOPIC: Machine Learning	TOPIC: Data Science	TOPIC: Visualization	TOPIC: Visualization
TUTORIAL	TUTORIAL	TUTORIAL	TUTORIAL
Pulling and Displaying ETF Data	Super Fast String Matching in Python	Understanding empirical Bayes estimation...	Brunel interactive visualizations in Jupyter...
AUTHOR: RStudio	AUTHOR: van den Blog	AUTHOR: Variance Explained	AUTHOR: Data Science Experience Blog
DATE: Feb 09, 2017	DATE: Nov 20, 2017	DATE: Mar 13, 2018	DATE: Jul 01, 2016
LEVEL: Intermediate	LEVEL: Intermediate	LEVEL: Intermediate	LEVEL: Intermediate

Tutorials

Search results (295) Sort by: Most Related

Popular filters: Spark Deep Learning Brunel

DATA SET	DATA SET	DATA SET	DATA SET
Adolescent fertility rate (births per 1,000...	Agriculture, value added (% of GDP) by...	Airbnb Data for Analytics: Amsterdam Calendar	Airbnb Data for Analytics: Amsterdam Listings
AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM
DATE: May 22, 2016	DATE: May 22, 2016	DATE: Dec 20, 2016	DATE: Dec 20, 2016
TOPIC: Society	TOPIC: Economy & Business	TOPIC: Economy & Business	TOPIC: Economy & Business
DATA SET	DATA SET	DATA SET	DATA SET
Airbnb Data for Analytics: Amsterdam Reviews	Airbnb Data for Analytics: Antwerp Calendar	Airbnb Data for Analytics: Antwerp Listings	Airbnb Data for Analytics: Antwerp Listings...
AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM	AUTHOR: IBM
DATE: Dec 20, 2016	DATE: Dec 20, 2016	DATE: Dec 20, 2016	DATE: Dec 20, 2016
TOPIC: Economy & Business	TOPIC: Economy & Business	TOPIC: Economy & Business	TOPIC: Business

Data Sets

# Watson Studio – Create Assets

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

IBM services

BigInsights HDFS	Cloud Object Storage	Cloud Object Storage (Infrastructure)	Cloudant
Compose for MySQL	Compose for PostgreSQL	DB2	DB2 for i
DB2 for z/OS	DB2 Hosted	DB2 on Cloud	DB2 Warehouse
Informix	Object Storage OpenStack Swift	Object Storage OpenStack Swift (Infrastructure)	PureData for Analytics
Watson Analytics			

Third-party services

Amazon Redshift	Amazon S3	Apache Hive	Cloudera Impala
Dropbox	Hortonworks HDFS	Microsoft Azure SQL Database	Microsoft SQL Server
MySQL	Oracle	Pivotal Greenplum	PostgreSQL
Remote file system transfer	Salesforce.com	Sybase	Sybase IQ
Teradata			

**Connect to Data Sources**

IBM Watson Projects Tools Catalog Community Services US South

My Projects / demo99 / Draw insights from Twitter da

File Edit View Insert Cell Kernel Help Not Trusted | Python 3.5

```

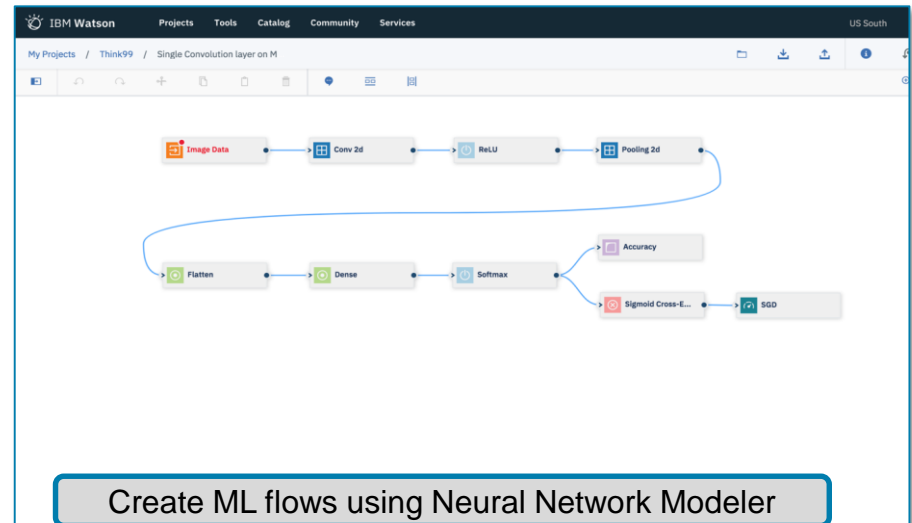
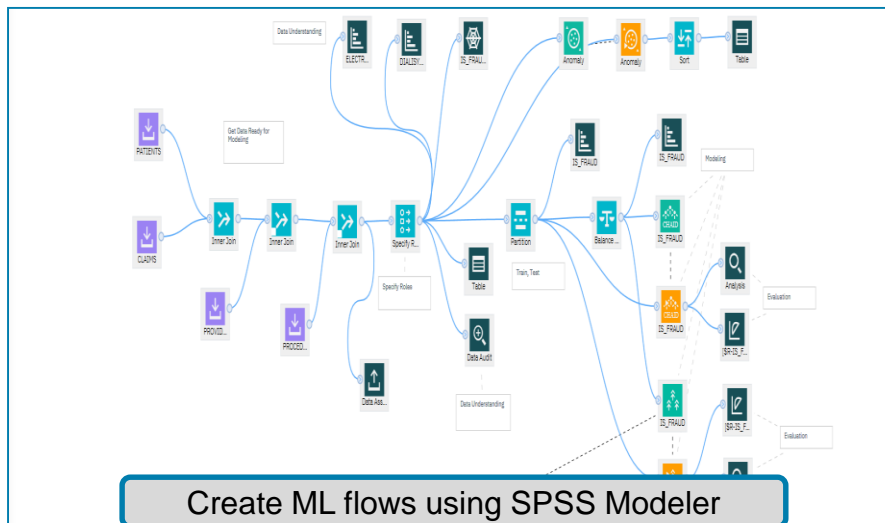
colors = ['gray'] + colors
it.figure(figsize=(10,8))
it.barh(y_pos, num_tweets, align='center', color=colors)
it.yticks(y_pos, countries)
it.xlabel('Number of Tweets')
it.title('Tweets Country Distribution based on the User Profile')
it.ylim(-1, len(y_pos))
it.show()

```

Tweets Country Distribution based on the User Profile

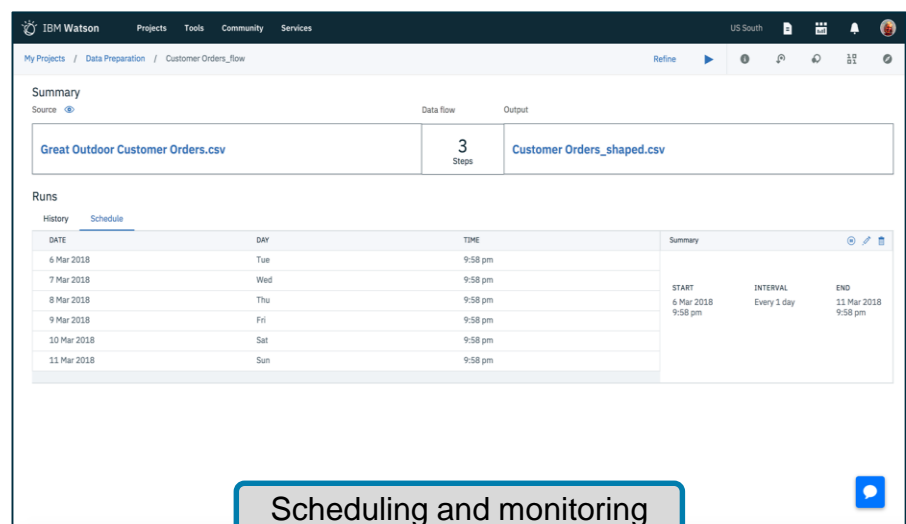
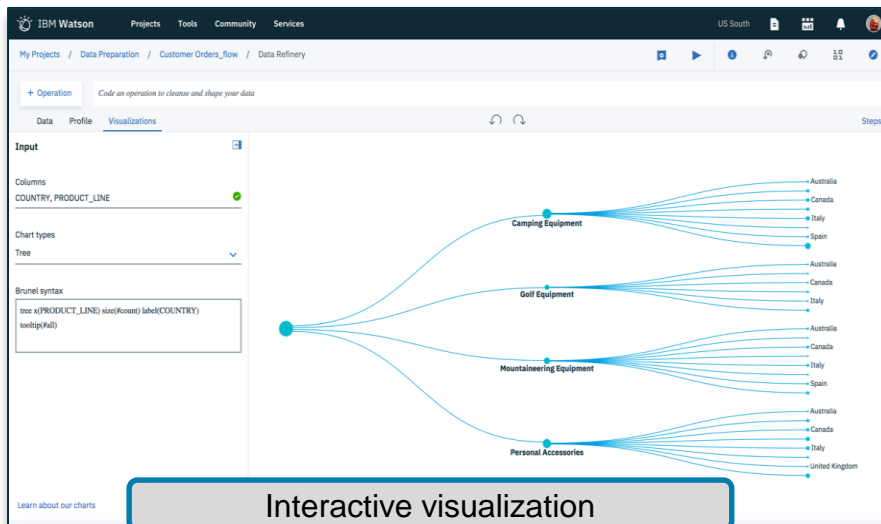
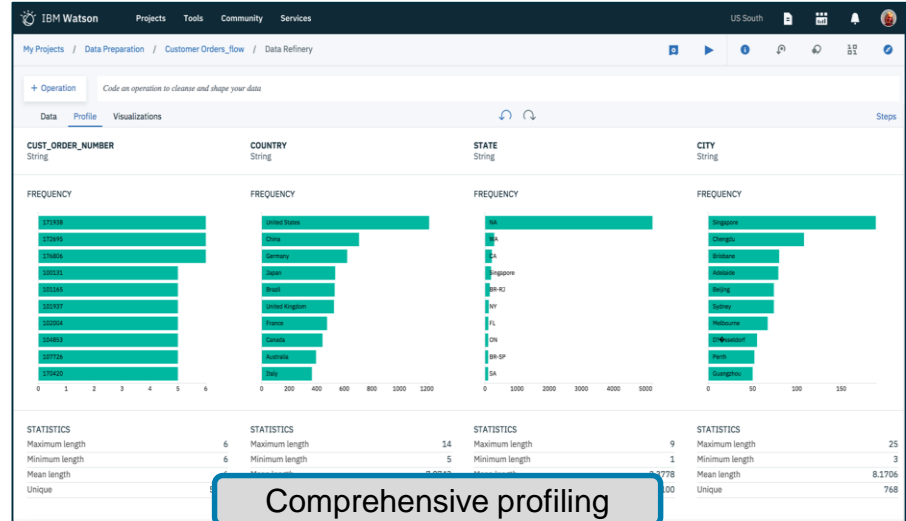
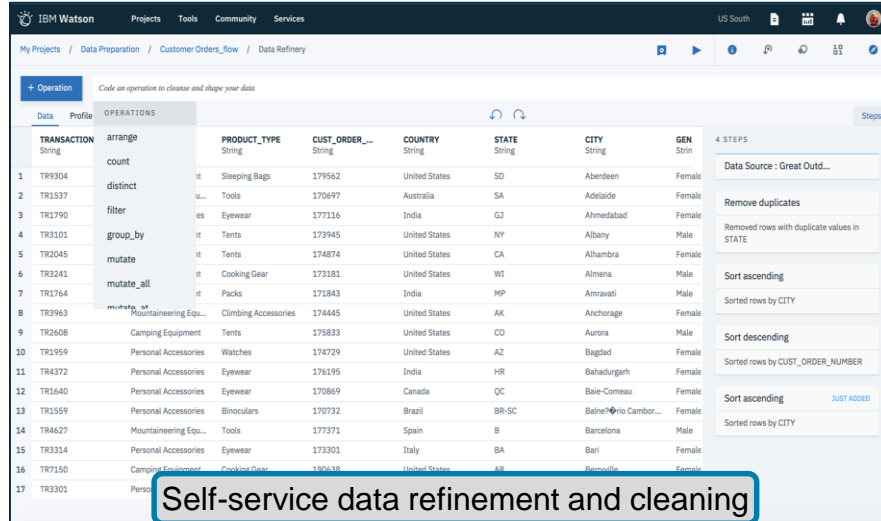
GERMANY  
MEXICO  
CANADA  
INDIA  
JAPAN  
SPAIN

**Open Source tools – Jupyter and RStudio**



# Watson Studio - Data Refinery

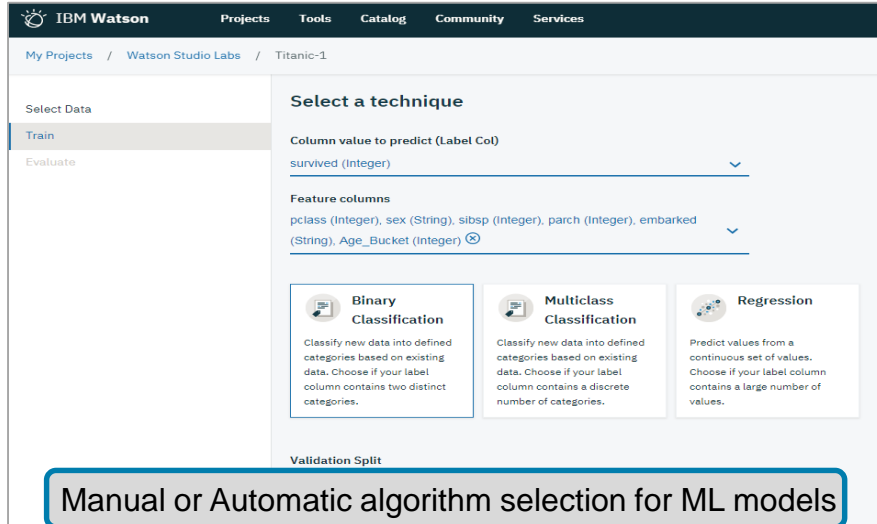
*Making Data fit for use*





# Watson Studio – Watson Machine Learning

*Simplifying deployment and management of ML models in production*



**Select a technique**

Column value to predict (Label Col)  
survived (integer)

Feature columns  
pclass (Integer), sex (String), sibsp (Integer), parch (Integer), embarked (String), Age\_Bucket (Integer)

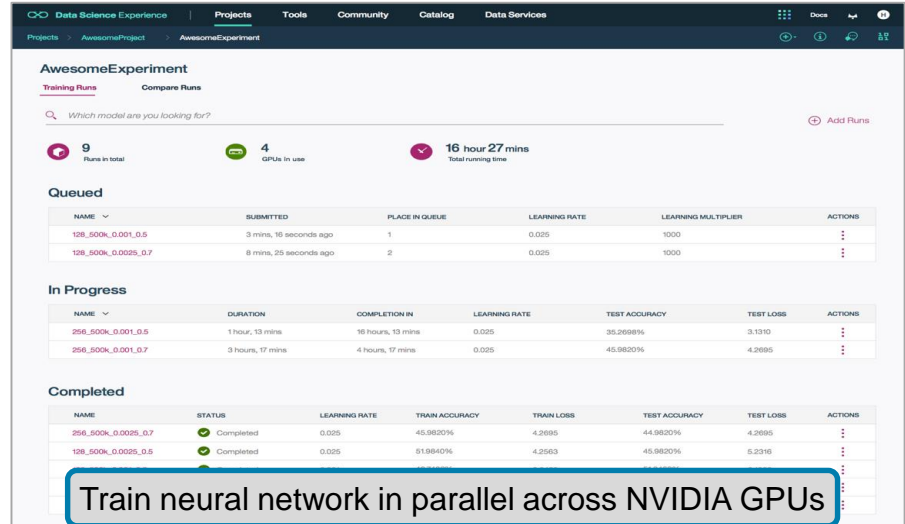
**Binary Classification**  
Classify new data into defined categories based on existing data. Choose if your label column contains two distinct categories.

**Multiclass Classification**  
Classify new data into defined categories based on existing data. Choose if your label column contains a discrete number of categories.

**Regression**  
Predict values from a continuous set of values. Choose if your label column contains a large number of values.

Validation Split

**Manual or Automatic algorithm selection for ML models**



**AwesomeExperiment**

Training Runs Compare Runs

Which model are you looking for?

9 Runs in total 4 GPUs in use 16 hour 27 mins Total running time

**Queued**

NAME	SUBMITTED	PLACE IN QUEUE	LEARNING RATE	LEARNING MULTIPLIER	ACTIONS
128_500k_0.001_0.5	3 mins, 16 seconds ago	1	0.025	1000	
128_500k_0.0025_0.7	8 mins, 25 seconds ago	2	0.025	1000	

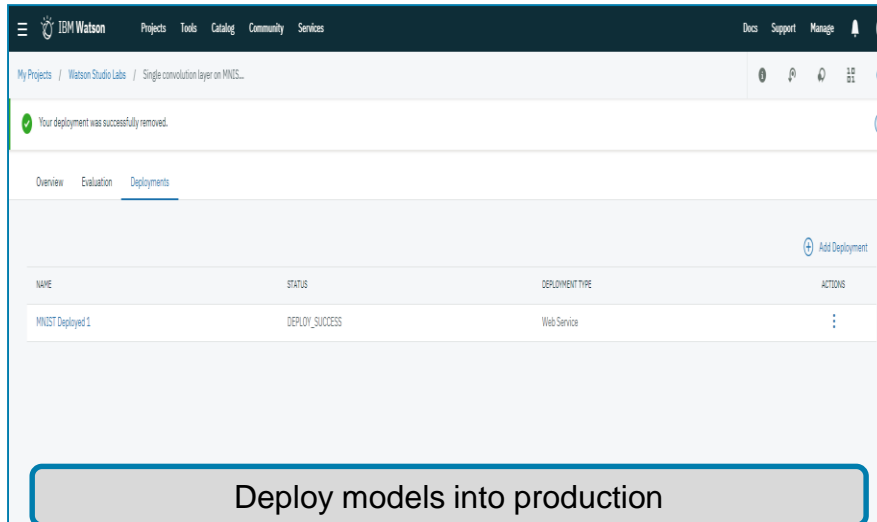
**In Progress**

NAME	DURATION	COMPLETION IN	LEARNING RATE	TEST ACCURACY	TEST LOSS	ACTIONS
256_500k_0.001_0.5	1 hour, 13 mins	16 hours, 13 mins	0.025	35.2688%	3.1310	
256_500k_0.001_0.7	3 hours, 17 mins	4 hours, 17 mins	0.025	45.9820%	4.2695	

**Completed**

NAME	STATUS	LEARNING RATE	TRAIN ACCURACY	TRAIN LOSS	TEST ACCURACY	TEST LOSS	ACTIONS
256_500k_0.0025_0.7	Completed	0.025	45.9820%	4.2695	44.9820%	4.2695	
128_500k_0.0025_0.5	Completed	0.025	51.9840%	4.2563	45.9820%	5.2316	

**Train neural network in parallel across NVIDIA GPUs**

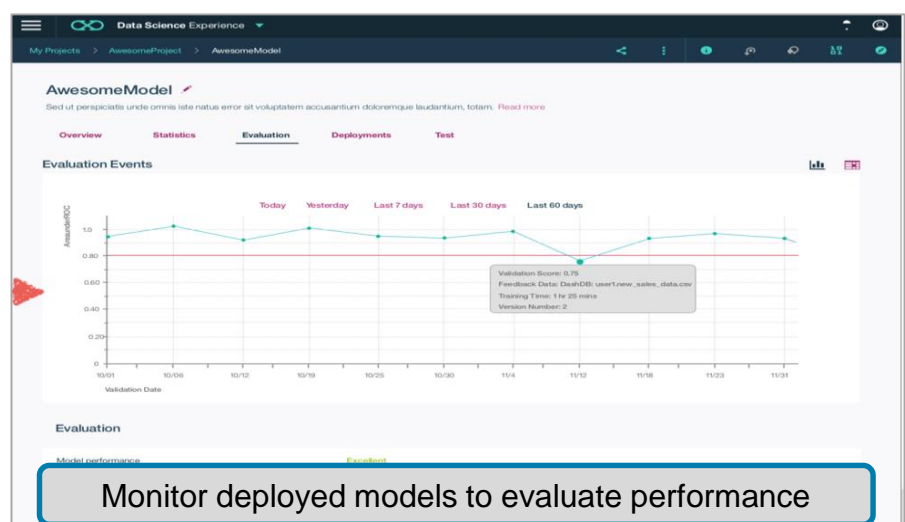


**Deployments**

NAME STATUS DEPLOYMENT TYPE ACTIONS

MNIST Deployed 1 DEPLOY\_SUCCESS Web Service

**Deploy models into production**



**AwesomeModel**

Overview Statistics Evaluation Deployments Test

**Evaluation Events**

Accuracy

Validation Date

Today Yesterday Last 7 days Last 30 days Last 60 days

Validation Score: 0.78  
Feedback Data: DashDB: user1new\_sales\_data.csv  
Training Time: 1 hr 25 mins  
Version Number: 2

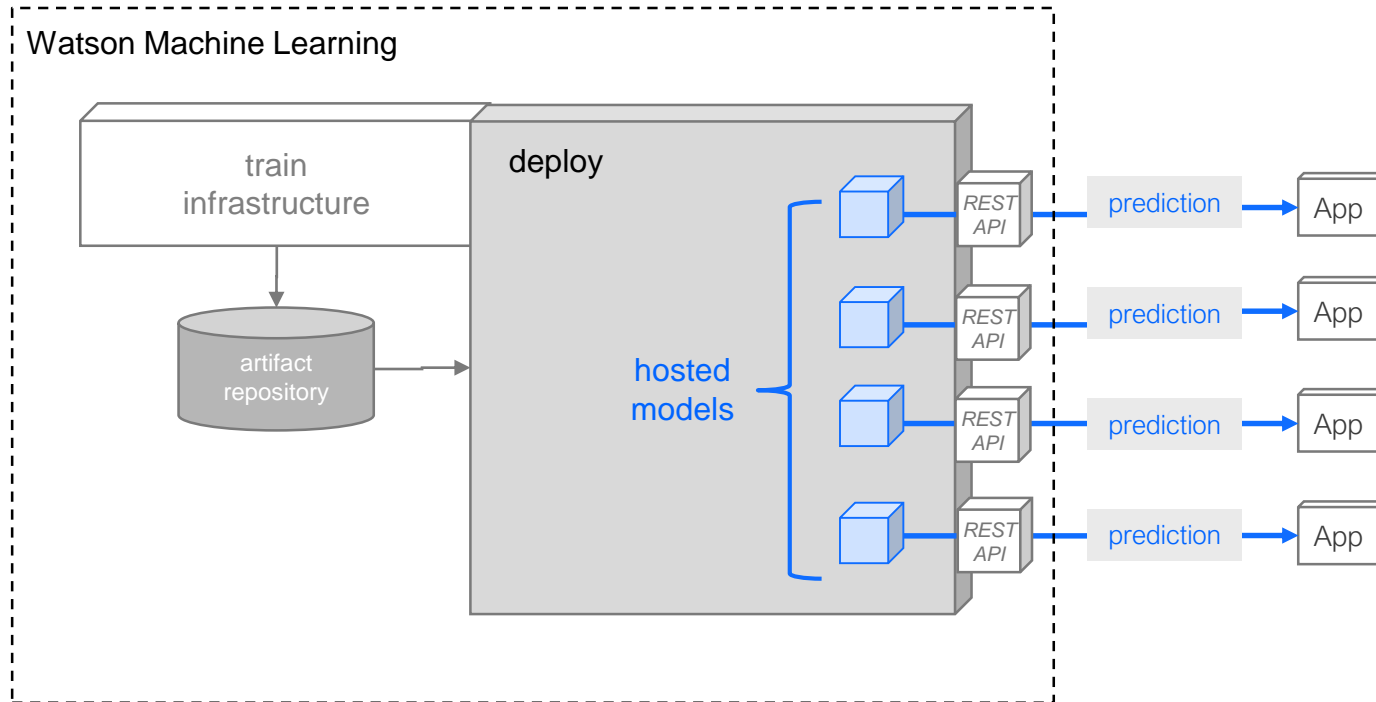
**Evaluation**

Model performance

**Monitor deployed models to evaluate performance**

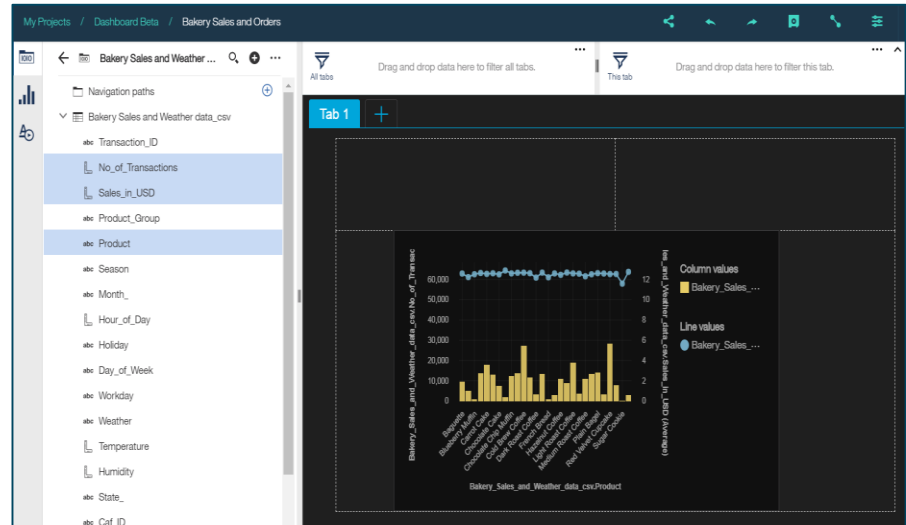
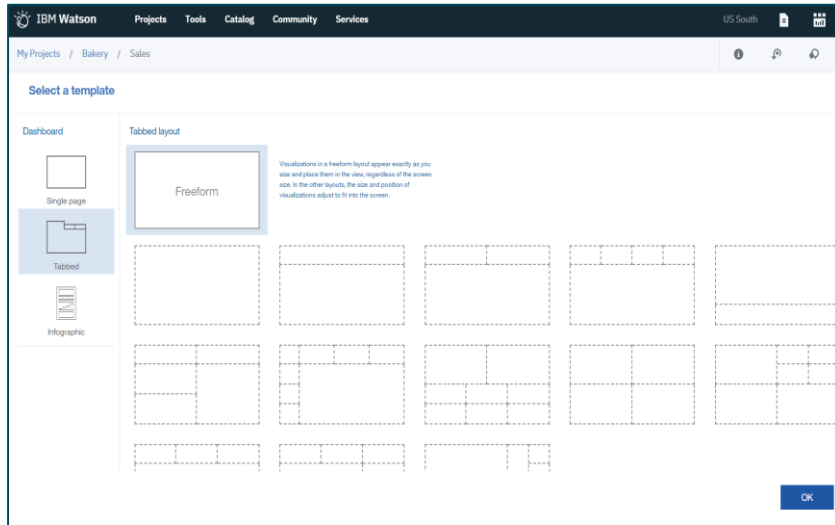
# Watson Studio- Deploying Trained Models

Deploy your models within Watson Machine Learning



# Watson Studio – Dynamic Dashboards

*Making insights available to all*



My Projects / Bakery Sales

**Data assets**

0 assets selected.

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

**Notebooks**

NAME SHARED SCHEDULED STATUS LANGUAGE LAST EDITOR LAST MODIFIED ACTIONS

Sales Predictions					Alex Jones	7 Mar 2018	
-------------------	--	--	--	--	------------	------------	--

**Streams flows**

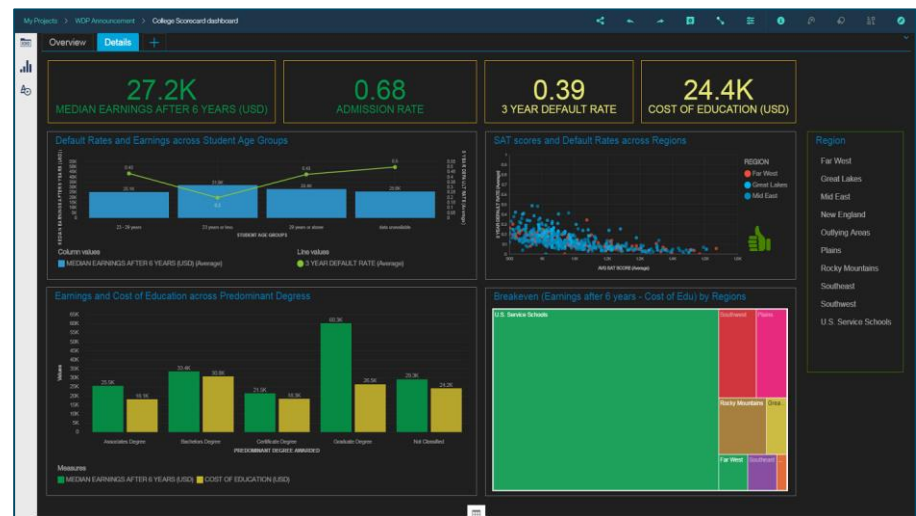
**Dashboard**

0 assets selected.

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

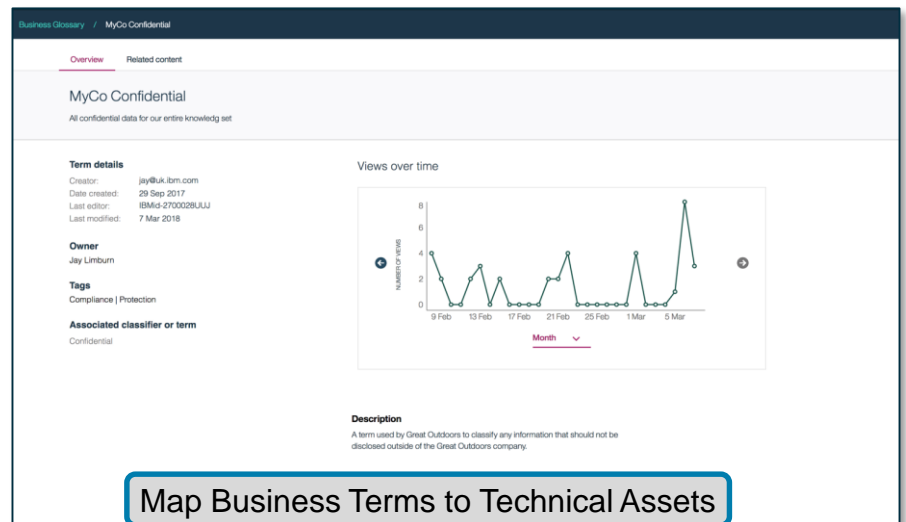
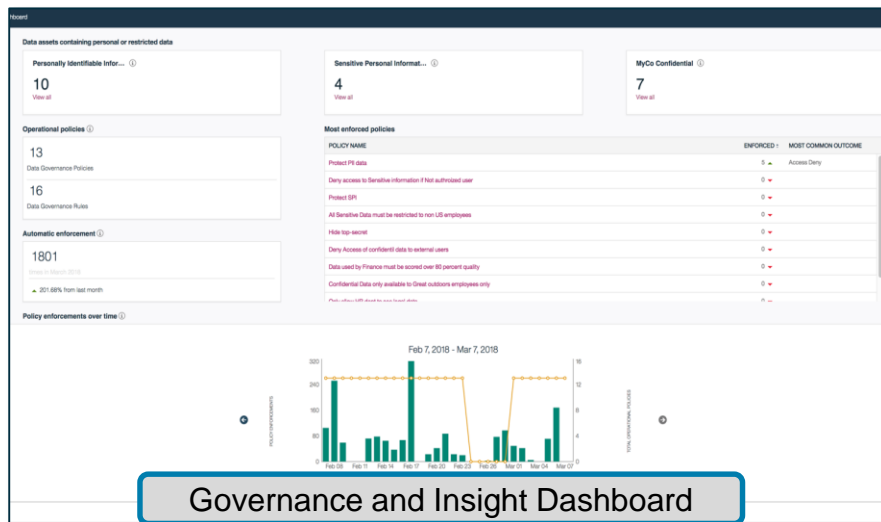
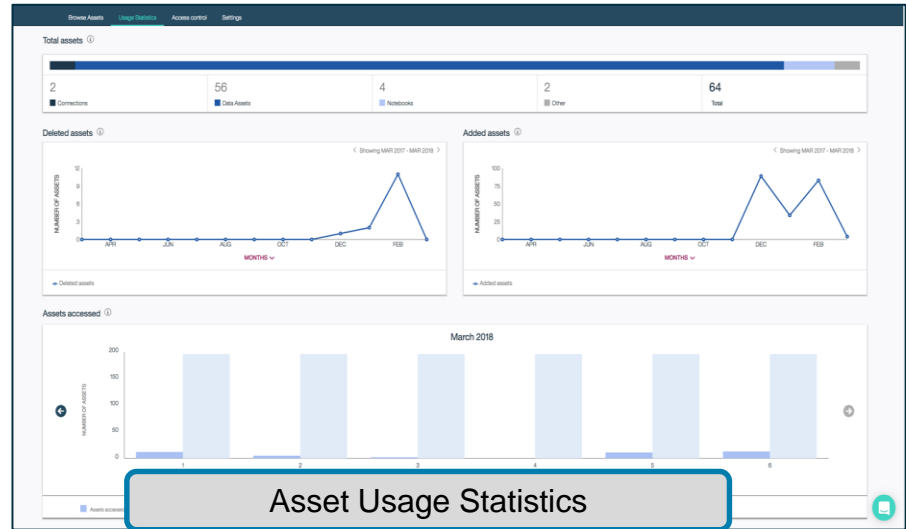
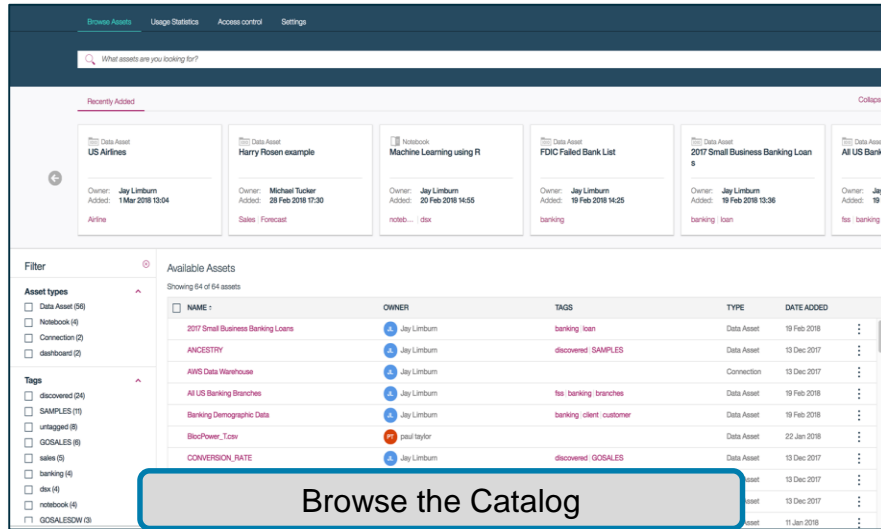
**Models**

Share Remove



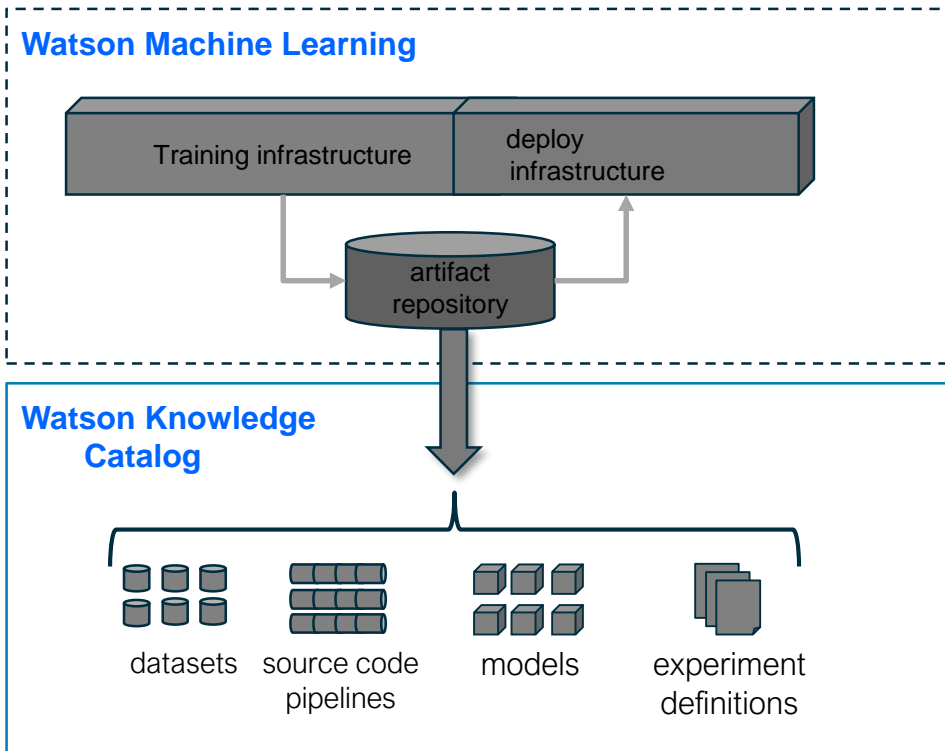
# Watson Knowledge Catalog

Unlock tribal knowledge and unleash knowledge workers



# 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 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

# 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](https://datascience.ibm.com)



# Lab Overview

## Lab Overview

Use IBM's Watson Studio and IBM cloud services to create a working cloud-based application from start to finish in Labs-1 and Lab-2. Use Watson Studio's Neural Network designer and Experiment Builder to build, train, and deploy a neural network model in Lab-3. Lab-4 will offer a choice of labs that explore additional features of Watson Studio.

- [Lab-1](#) - The first lab will leverage Spark machine learning (SparkML) in a Jupyter notebook to create categorical predictions using pyspark and a supervised learning model. The model will be saved into a model repository using Watson Machine Learning APIs.
- [Lab-2](#) - The second lab will guide participants in examining an R notebook and Shiny UI in Watson Studio using RStudio. It will rely on the output results from Lab-1.
- [Lab-3](#) - The third lab will use IBM's Neural Network designer and Experiment builder to build and train a simple Convolutional Neural Network. We will use the well-known MNIST dataset for training, test, and validation.
- [Lab-4](#) - Time permitting there will be three labs to choose from for Lab 4. The first one features Watson Machine Learning, a point and click capability to build a machine learning model and deploy it. The second lab features the SPSS Modeler - a visual programming tool to create a machine learning pipeline. The third lab features the Data Refinery tool a fully managed self-service data preparation facility.

## Lab Tips

- Labs are all located in [www.github.com/bleonardb3/WatsonStudio](https://www.github.com/bleonardb3/WatsonStudio) repository. Environment set up is located in the repository [README](#) file.
- 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-3 and Labs-4a,4b, and 4c. 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 1, you execute notebook cells by <Shift><Enter> when your cursor is in a code cell.
- For Lab 1, you will see instructions to use and Insert to Code facility. Make sure the notebook cursor is in the correct cell prior to doing the insert.

## Lab 1

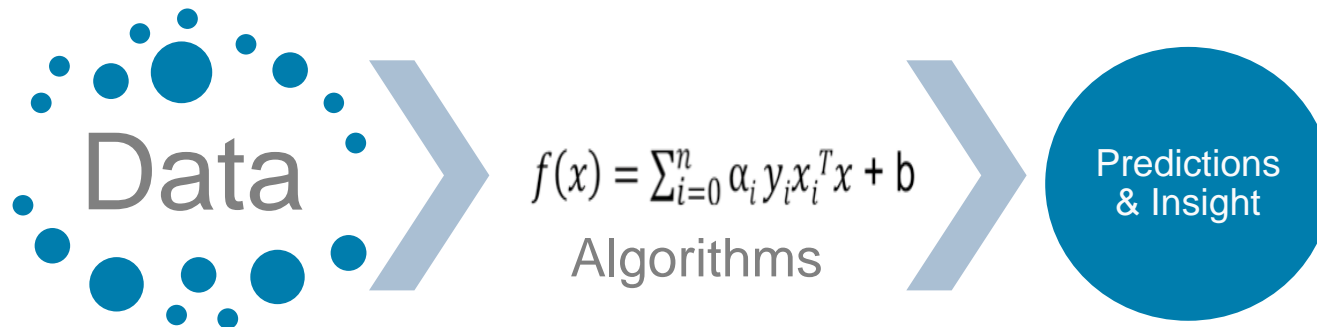
In this lab, you will use SparkML in Watson Studio to run generated travel data through a machine learning algorithm, automatically tune the algorithm, and load the data into Cloud Object Storage.

### Objectives:

- Upon completing the lab, you will know how to:
  - Connect to cloud object storage and read data used for machine learning.
  - Identify labels and transform data.
  - Conduct feature engineering for algorithm data.
  - Declare a machine learning model.
  - Setup the Pipeline for data transforms and training.
  - Train the data.
  - Show and evaluate machine learning results.
  - Automatically tune machine learning results.
  - Score data and load results into cloud object storage.

# What is Machine Learning?

*“Computers that learn without being **explicitly programmed**”*  
*“Using **algorithms** to understand patterns in data”*



# 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

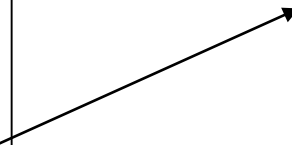
# Categories of Machine Learning

Technique	Usage	Algorithms
Classification (or prediction)	<ul style="list-style-type: none"><li>• Used to predict group membership (e.g., will this employee leave?) or a number (e.g., how many widgets will I sell?)</li></ul>	<ul style="list-style-type: none"><li>• Decision Trees</li><li>• Logistic Regression</li><li>• Random Forests</li><li>• <b>Naïve Bayes</b></li><li>• Linear Regression</li><li>• Lasso Regression</li><li>etc</li></ul>
Segmentation	<ul style="list-style-type: none"><li>• Used to classify data points into groups that are internally homogenous and externally heterogeneous.</li><li>• Identify cases that are unusual</li></ul>	<ul style="list-style-type: none"><li>• K-means</li><li>• Gaussian Mixture</li><li>• Latent Dirichlet allocation</li><li>etc</li></ul>
Association	<ul style="list-style-type: none"><li>• Used to find events that occur together or in a sequence (e.g., market basket)</li></ul>	<ul style="list-style-type: none"><li>• FP Growth</li></ul>

Known as:

- Scale variables:

- Categorical variables:

- | a1  | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 | t |
|---|----|----|----|----|----|----|----|----|---|
|  |    |    |    |    |    |    |    |    |   |

- 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

# 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

# Lab 1 – Female Human Trafficking


## ▪ Input


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

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

# Lab 1 Data

Field	Description
UUID	Hash-based unique identifier
<b>VETTING_LEVEL</b>	Analyst vetting status : 100- PENDING, 10 – HIGH, 20 – MED, 10 - LOW
NAME	Person name
<b>GENDER</b>	Person Gender
AGE	Person age at time of travel
<b>BIRTH_DATE</b>	Person birth date
BIRTH_COUNTRY	Person full birth country
BIRTH_COUNTRY_CODE	Person ISO 2 country
<b>OCCUPATION</b>	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
<b>PASSPORT_COUNTRY_CODE</b>	Person Passport Issuing Country ISO 2 Code
COUNTRYIES_VISITED	The countries visited as declared on form
<b>COUNTRIES_VISITED_COUNT</b>	The number of countries visited as declared on form
ARRIVAL_AIRPORT_COUNTRY_CODE	ARRIVAL Airport country code ISO2
AIRPORT_ARRIVAL_IATA	ARRIVAL Airport 3 character code
AIRPORT_ARRIVAL_MUNICIPALITY	ARRIVAL Airport Municipality Derived from Code
ARRIVAL_AIRPORT_REGION	ARRIVAL Airport Region Derived from Code
DEPARTURE_AIRPORT_COUNTRY_CODE	DEPARTURE Airport Country code ISO2
DEPARTURE_AIRPORT_IATA	DEPARTURE Airport 3 character code
DEPARTURE_AIRPORT_MUNICIPALITY	DEPARTURE Airport Municipality Derived from Code.

 Target

 Features

## Lab 1 Flow

- **Read in dataset from Cloud Object Storage**
  - Connect to Object Storage
  - Read in the 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
- **Train the Model**
  - Split input data into Training (80%) and Test (20%) DataFrames
  - Cache the resulting DataFrames
  - Fit the Pipeline to the Training data set



## Lab 1 Flow (continued)

- **Evaluate the resulting predictions**
  - Area under the ROC curve
  
- **Tune the model (hyperparameters)**
  - 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 Cloud Object Storage
  
- **Save the model in the Model Repository**
  - Model properties can be saved as well (e.g Area under the ROC curve)

## Lab 2 – RStudio and Shiny

In this lab, you will learn some of the fundamentals of using RStudio and Shiny in Watson Studio to work and interact with data and then to create a fully operational "reactive" web application that you can enhance further.

### Objectives:

- Upon completing the lab, you will know how to:
  - Create an RStudio project from a Git repository
  - Establish a connection to Cloud Object Storage using an ancillary file
  - Query, join, explore and visualize data in an R notebook
  - Derive categorical names from numerical levels in an R dataframe
  - Use ggplot2 to create bar plots of several of the columns in an R dataframe
  - Use a logarithmic scale when creating bar plots
  - Leverage shiny to create and run a web application
  - Interact with the shiny web application by running it externally



## Lab 3 – Neural Network

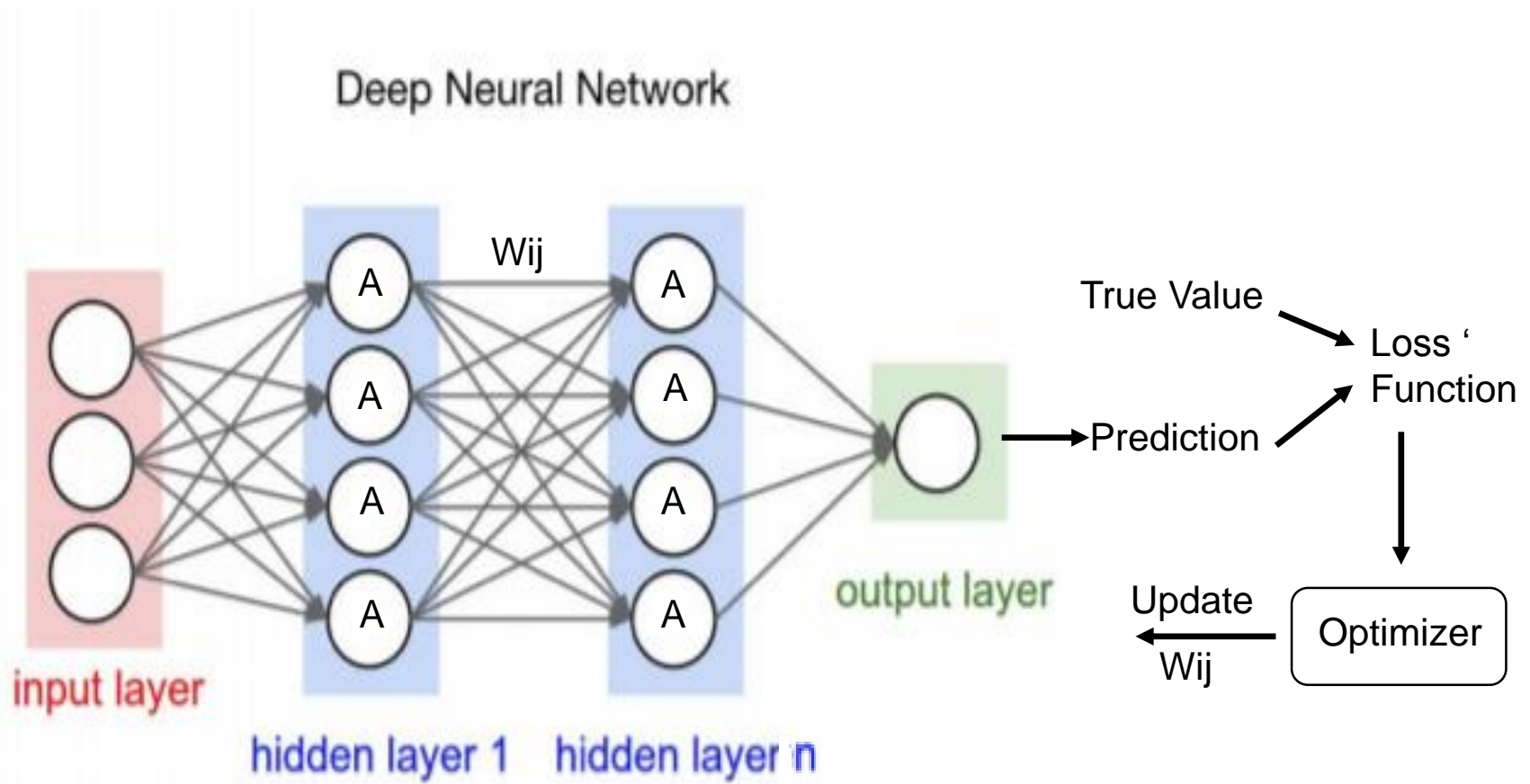
This lab will use the [MNIST](#) computer vision data set to train a convolutional neural network (CNN) model to recognize handwritten digits. The Watson Studio neural network flow editor, Watson Studio experiment builder and the Watson Machine Learning component will be used to build, train, and save the trained model.

### Objectives:

- Upon completing the lab, you will know how to:
  - Create Cloud Object Storage buckets to contain the input and result files
  - Create a neural network design from an example using the flow editor
  - Use the experiment builder used to set up a training definition to train the neural network model
  - Monitor the training progress and results.
  - Save the trained model.

# Neural Network

- Inspired by the way the human brain works.



$W_{ij}$  - weights

A – Activation Function

# Neural Network

Modeling

- **Originated in 1940s**
- **Became very popular this decade**
  - Hardware – GPUs, Storage
  - Availability of Large Datasets for Training
  - Better performing algorithms.
- **Especially useful for human perception type task**
  - Image Classification
  - Object Recognition
  - Speech Recognition
  - Natural Language Understanding
  - Machine Translation
  - ...

# Neural Network Modeler

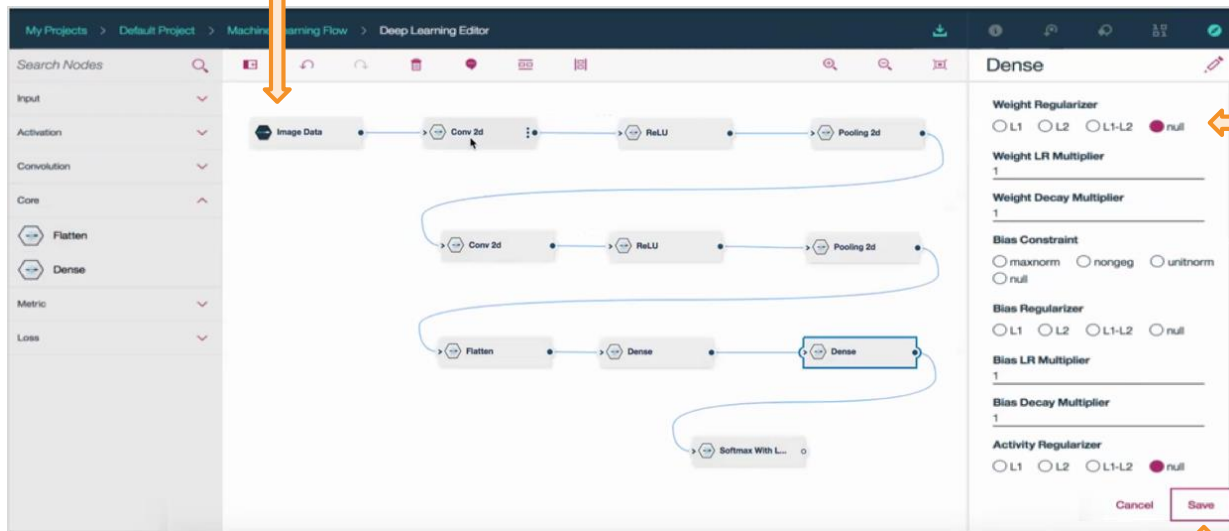
An intuitive drag-and-drop, no-code interface for designing neural network structures using the most popular deep learning frameworks. Quickly capture your network design then single click export for experimental optimization.

## Supported Frameworks



Drag-and-drop  
network layers

Real-time validation of network  
flow

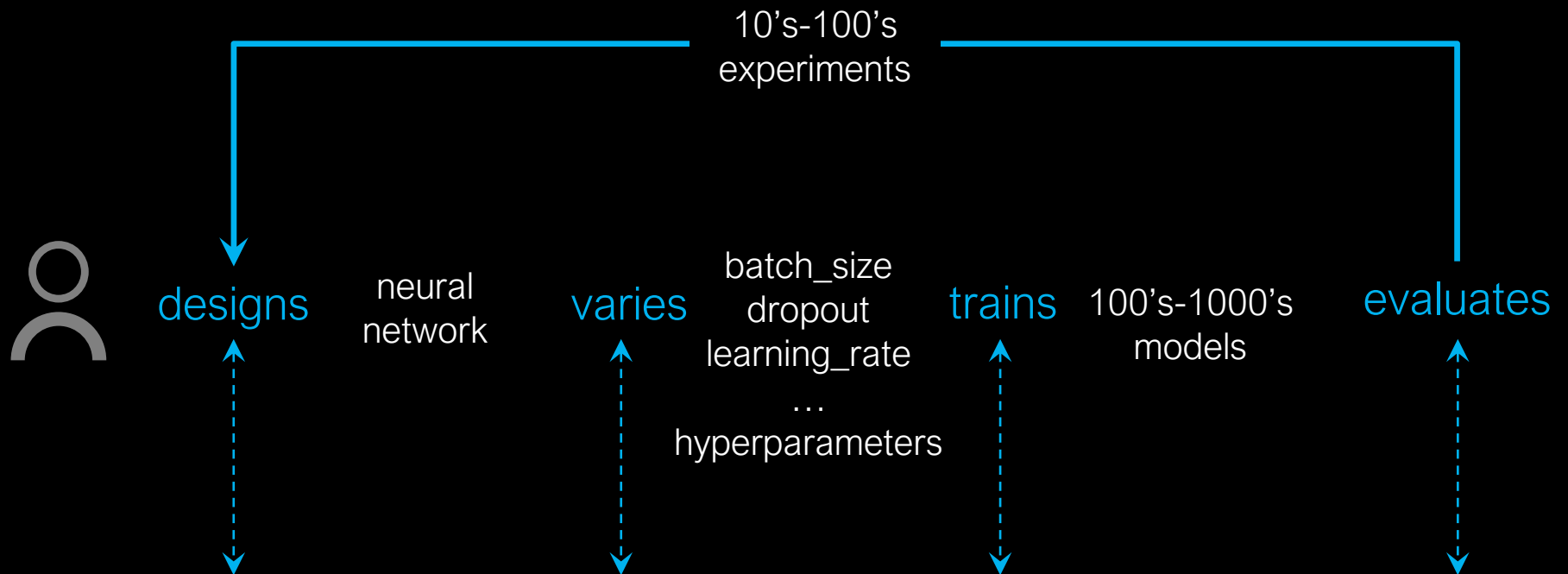


- Define layer configuration
- Choose optimizer params

- Generate CPU or GPU compatible code

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

# Experiment Builder



Experiment Builder  
supports the end-to-end workflow

## Lab 4a – 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.
  - Deploy an application that invokes the machine learning model service.

## Lab 4b – 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 4c – 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.



# 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