

Dyphen Vian

Workshop de Machine Learning

Desafío DUOC CITT 2020



ROBERTO MUÑOZ
SENIOR DATA SCIENTIST
EY METRICARTS



The better the question. The better the answer. The better the world works.



EY

Building a better
working world

Analytics

Uso intensivo de datos, estadística y modelos predictivos



1968

1973

1974

1981

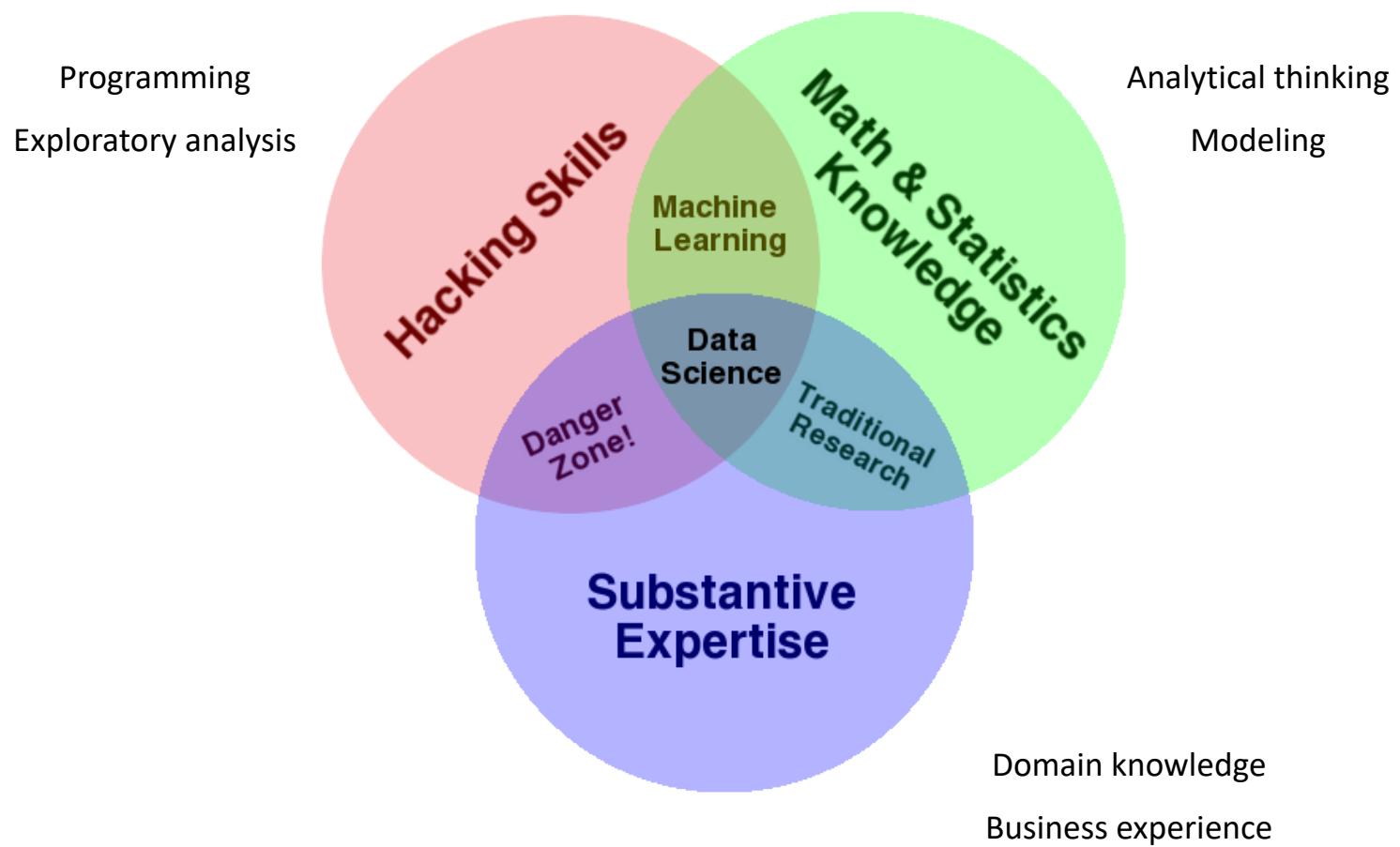


2015

2006

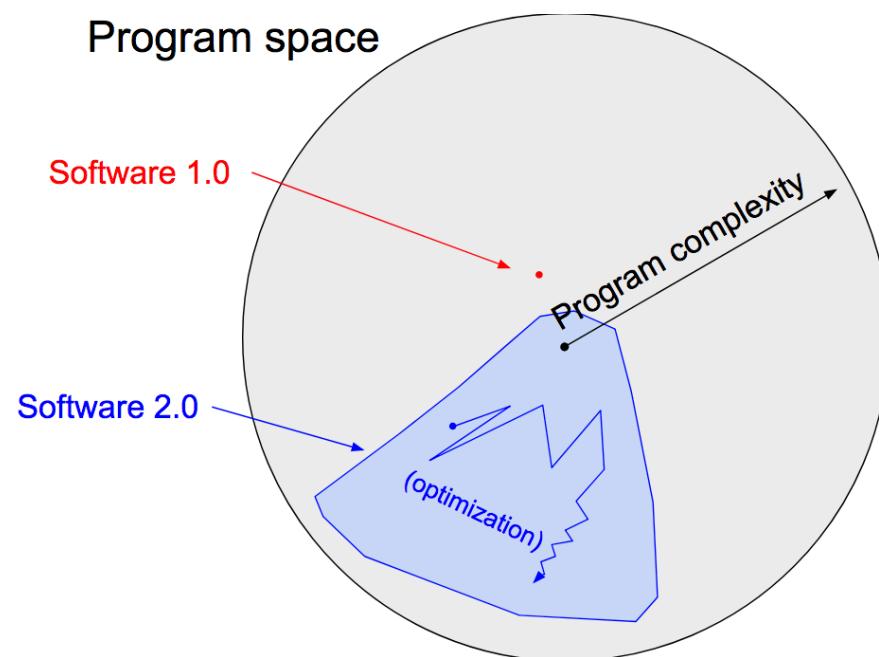
1993

1991



The Data Science Venn Diagram
Drew Conway (2010)

Software 2.0



Software 1.0 is what we're all familiar with — it is written in languages such as Python, C++, etc. It consists of explicit instructions to the computer written by a programmer

Software 2.0 can be written in much more abstract, human unfriendly language, such as the weights of a neural network. No human is involved in writing this code because there are a lot of weights and coding directly in weights is kind of hard

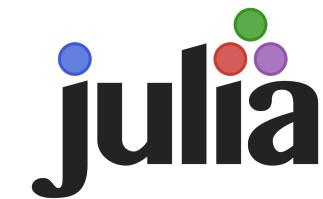
Andrej Karpathy

Director of AI at Tesla

Programming languages

The most used languages in Data Science are Python and R. Python has more than 30 million (M) users and R more than 16M.

Julia has emerged as an all-around and efficient language. Around 2M users.



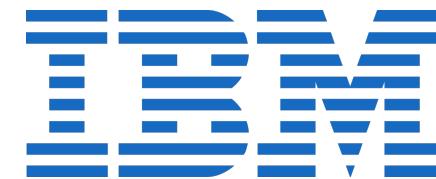
Infrastructure



Google
Cloud Platform



Microsoft
Azure





Machine Learning

Cuando debería usar Machine learning

Tipos de ML

Aprendizaje
supervisado



Regresión: ¿Cuál será la temperatura dentro de 5 horas?

Clasificación: ¿A qué clase pertenece esta imagen/texto?

Aprendizaje no
supervisado



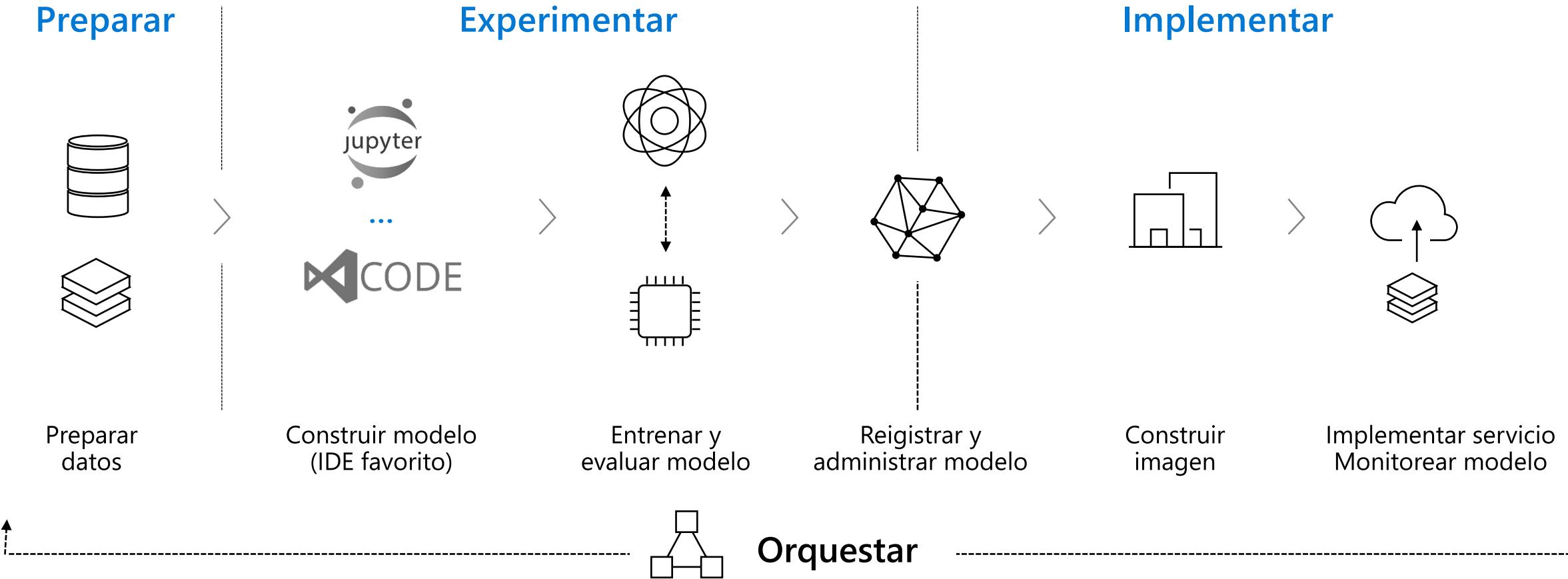
Clustering: ¿Se pueden agrupar los datos? ¿A qué grupo pertenece?

Detección de anomalías: ¿Es este valor extraño?

Recomendación: ¿Qué opción debería elegir?

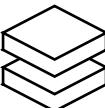
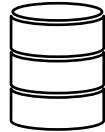
Machine Learning

Proceso end-to-end

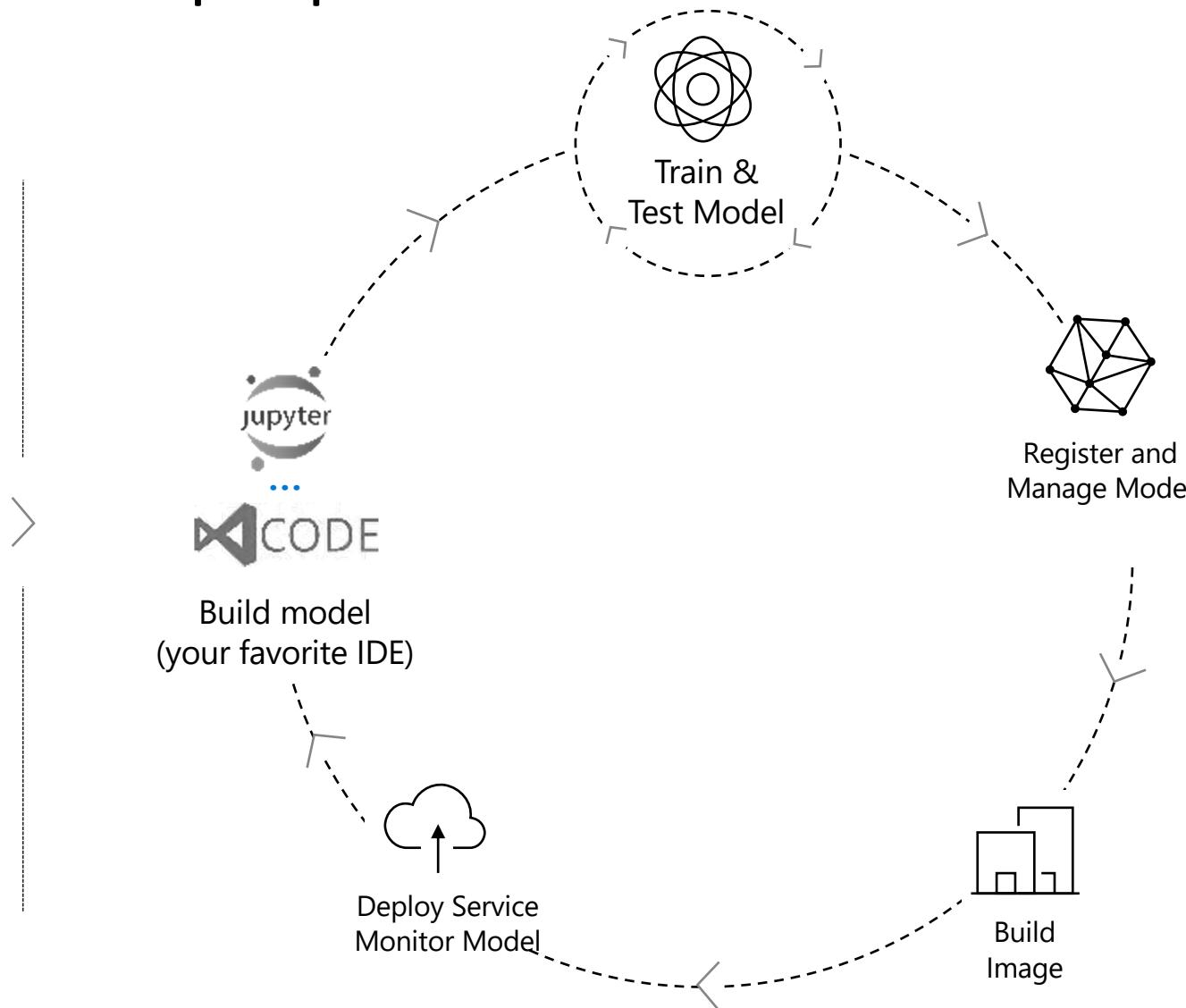


Ciclo DevOps para Data science

Preparar



Prepare
Data



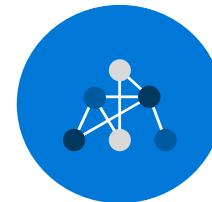
Azure AI

Apps y agentes de AI



Azure Bot Service
Azure Cognitive Services

Machine learning



Azure Machine Learning

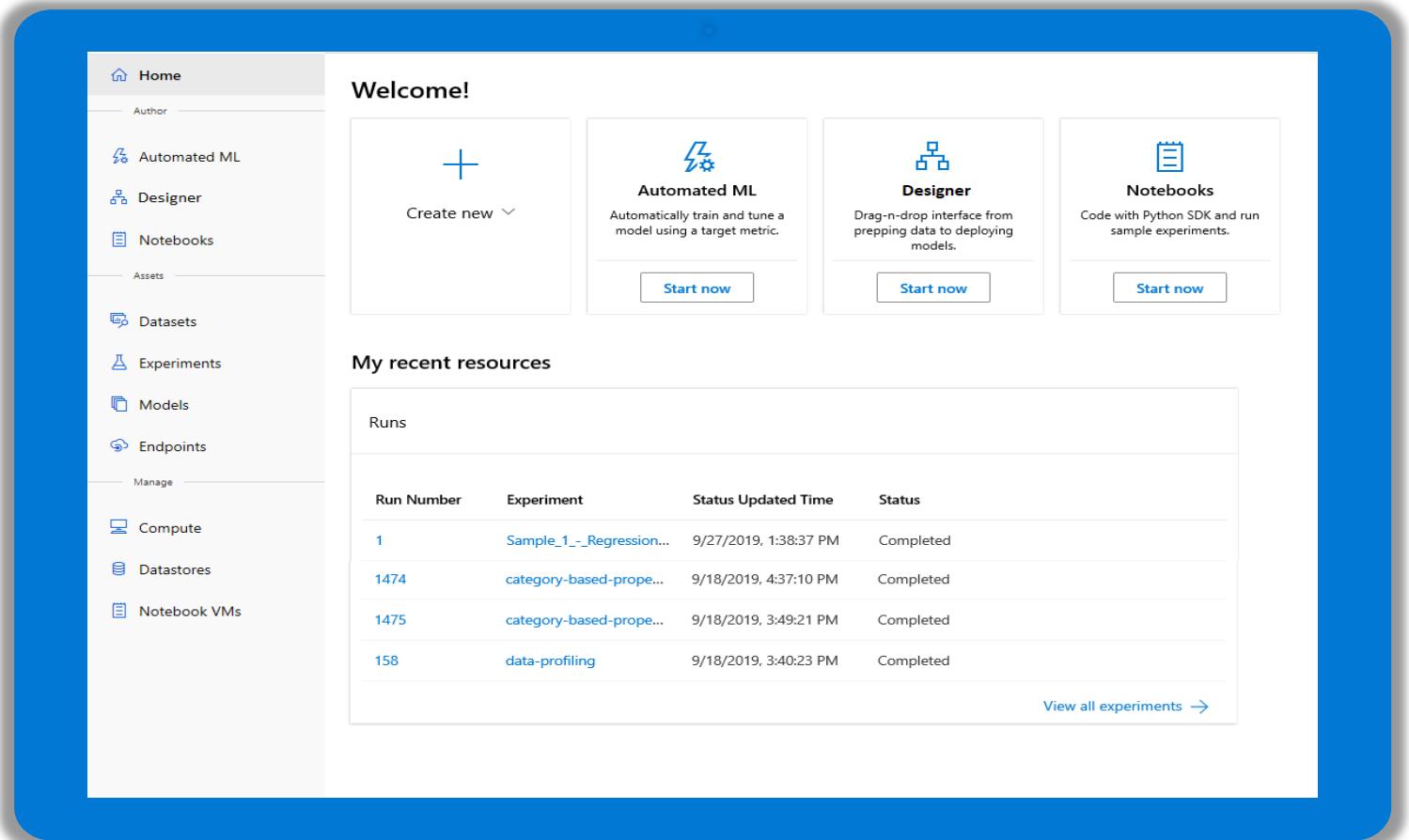
Minería de conocimiento



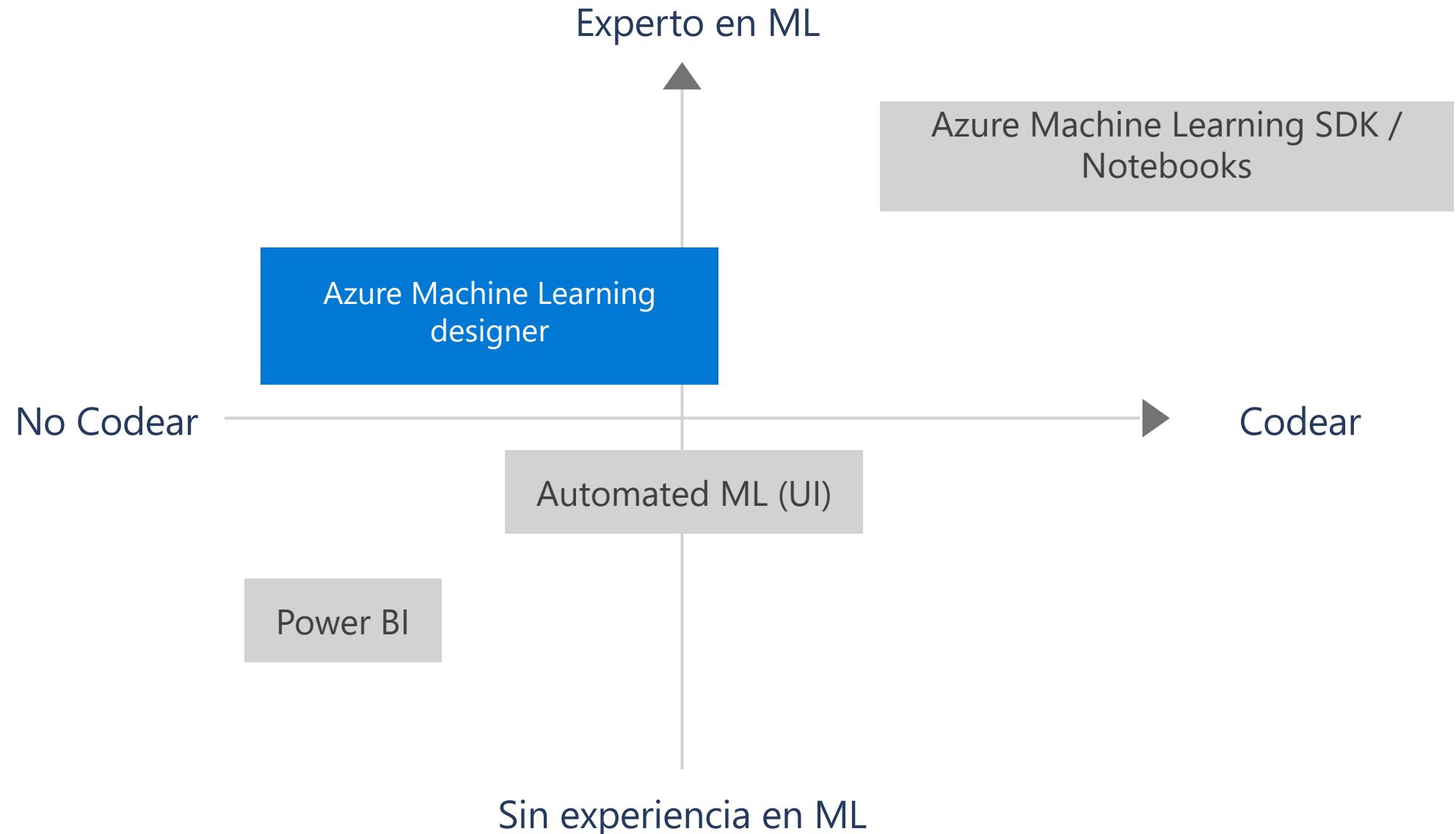
Azure Cognitive Search

Azure Machine Learning

Para todos los niveles



Build across skill levels with Azure Machine Learning

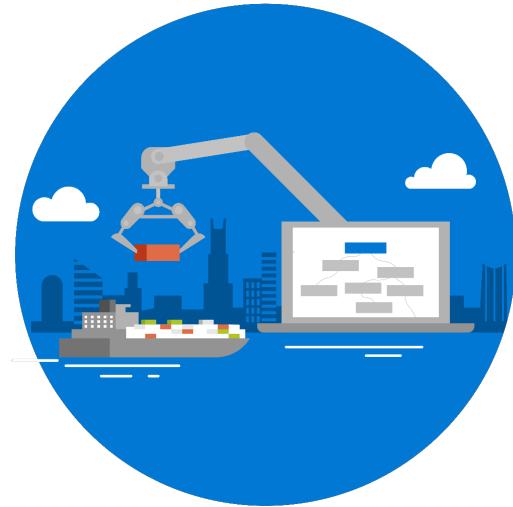


¿Quienes usan el Desginer de Azure ML?



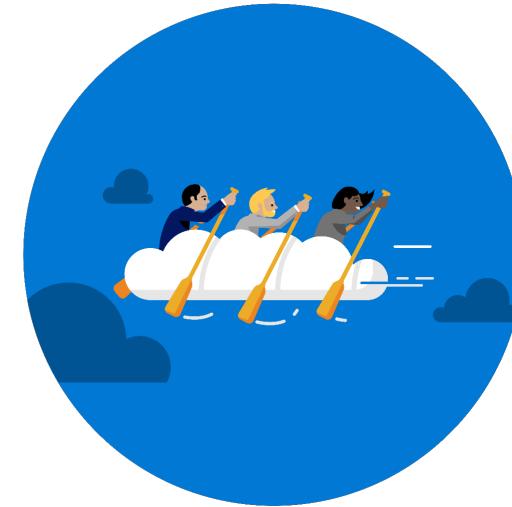
Neófito en Machine Learning

Aprender machine learning



Data Scientists junior

Construir modelos y preferir herramientas visuales vs programar



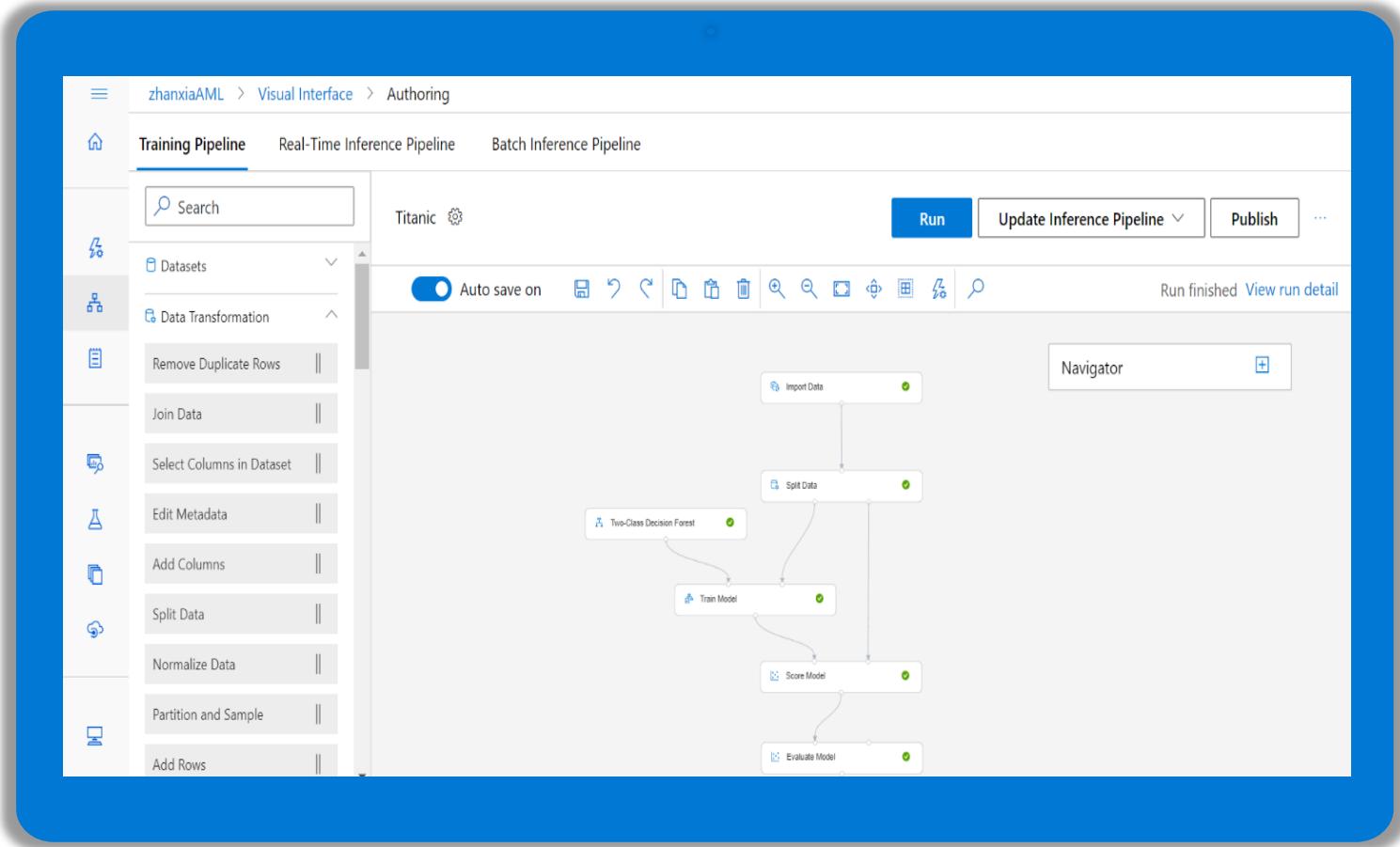
Expertos en ML

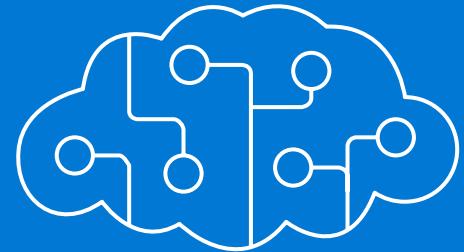
Prototipado rápido



Azure Machine Learning designer

- Simplifique el proceso de creación, prueba y funcionamiento de modelos de aprendizaje automático.
- Módulos y ejemplos integrados para comenzar y crear modelos personalizados
- Soporte para R y Python
- Comparta fácilmente entre espacios de trabajo





Azure Machine Learning: Technical Details

Azure ML service

Key Artifacts



Workspace



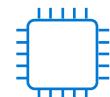
Models



Experiments



Pipelines



Compute Target



Images



Deployment



Data Stores

Azure ML service Artifacts

Models and Model Registry



Model

A machine learning model is an artifact that is created by your training process. You use a model to get predictions on new data.

A model is produced by a **run** in Azure Machine Learning.

Note: You can also use a model trained outside of Azure Machine Learning.

Azure Machine Learning service is framework agnostic — you can use any popular machine learning framework when creating a model.

A model can be registered under an Azure Machine Learning service workspace



Model Registry

Keeps track of all the models in your Azure Machine Learning service workspace.

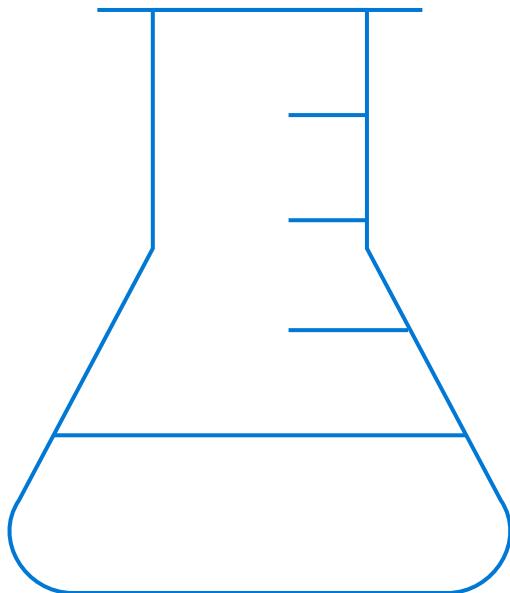
Models are identified by name and version.

You can provide additional metadata tags when you register the model, and then use these tags when searching for models.

You cannot delete models that are being used by an image.

Azure ML Artifacts

Runs and Experiments



Experiment

Grouping of many runs from a given script.

Always belongs to a workspace.

Stores information about runs

Run

Produced when you submit a script to train a model. Contains:

Metadata about the run (timestamp, duration etc.)

Metrics logged by your script.

Output files autocollected by the experiment, or explicitly uploaded by you.

A snapshot of the directory that contains your scripts, prior to the run.

Run configuration

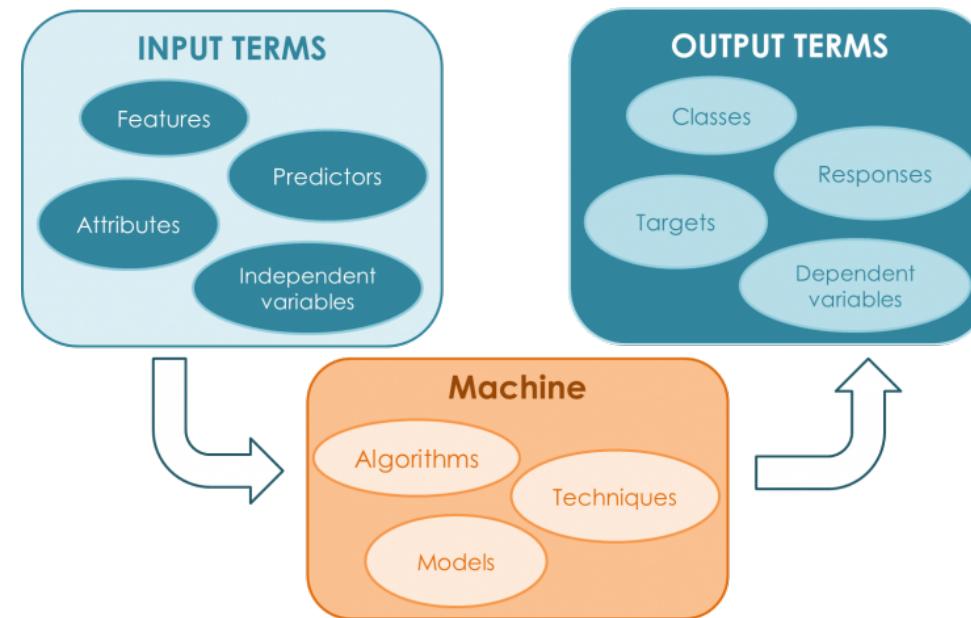
A set of instructions that defines how a script should be run in a given compute target.

Machine Learning



Machine learning

- *Machine learning* is a subfield from Computer Science and Statistics. Heavily based in programming and statistics
- The goal is to develop techniques that allow to computers to learn or imitate human cognitive skills



Usual tasks

- **Regression**

“The company Entel wants to learn how many GB their clients will consume next month”

- **Classification**

“Falabella wants to promote a new product. What gender and age they should focus their ads?”

- **Clustering**

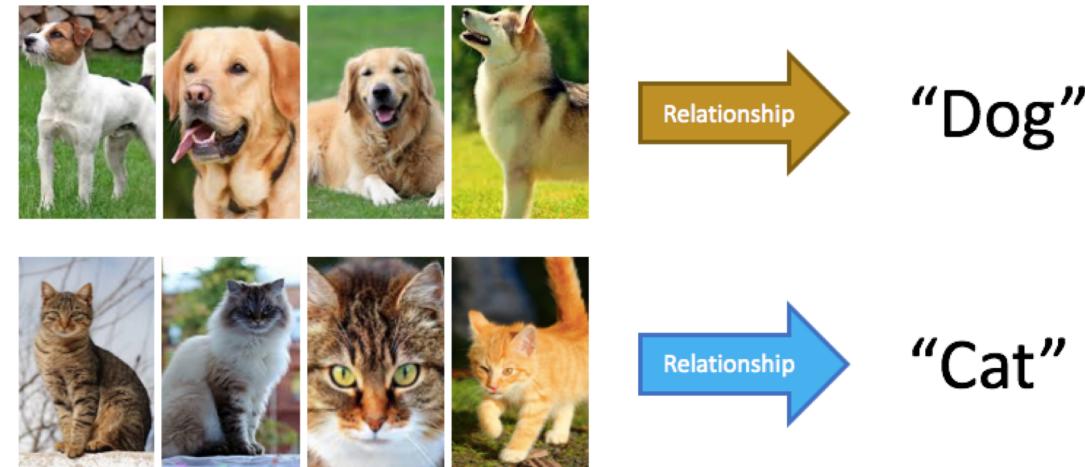
“The ESO observatories want to identify anomalous images produced by their telescopes. They don’t have labeled data.”

Types of ML

- Supervised learning

The system learns using previously classified data. The data can be structured or unstructured.

Algorithm generates a model that establish the correspondence between the input data and expected output of the system.

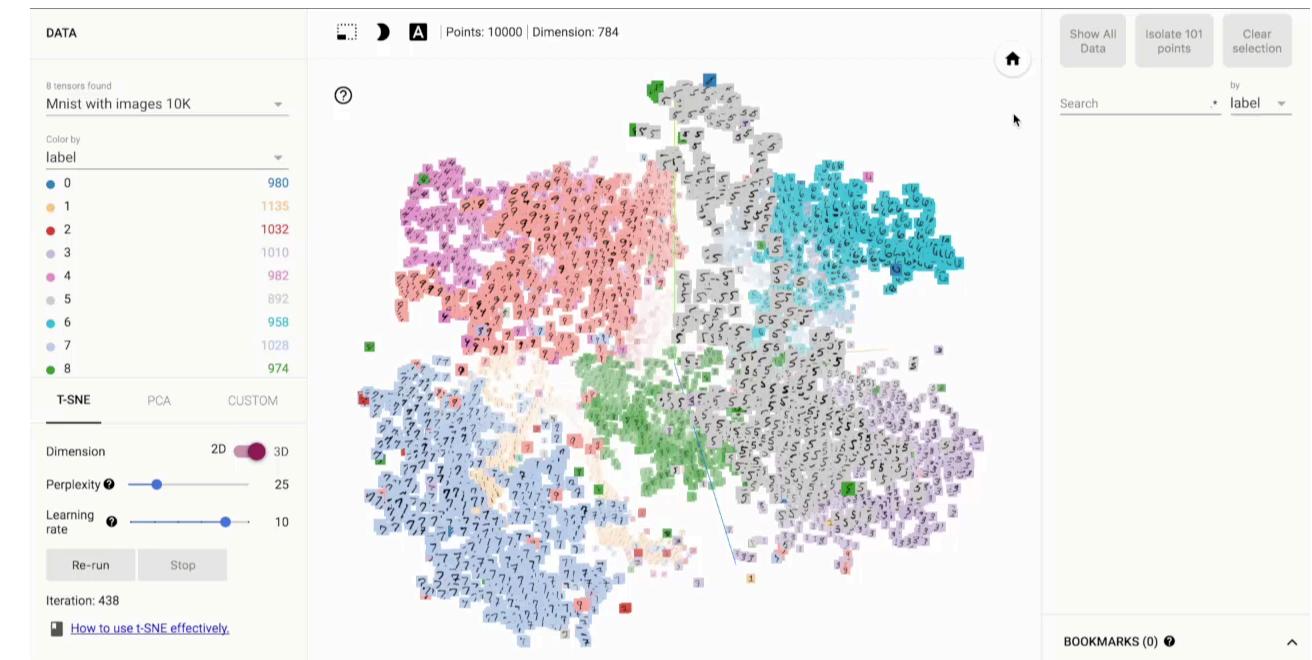
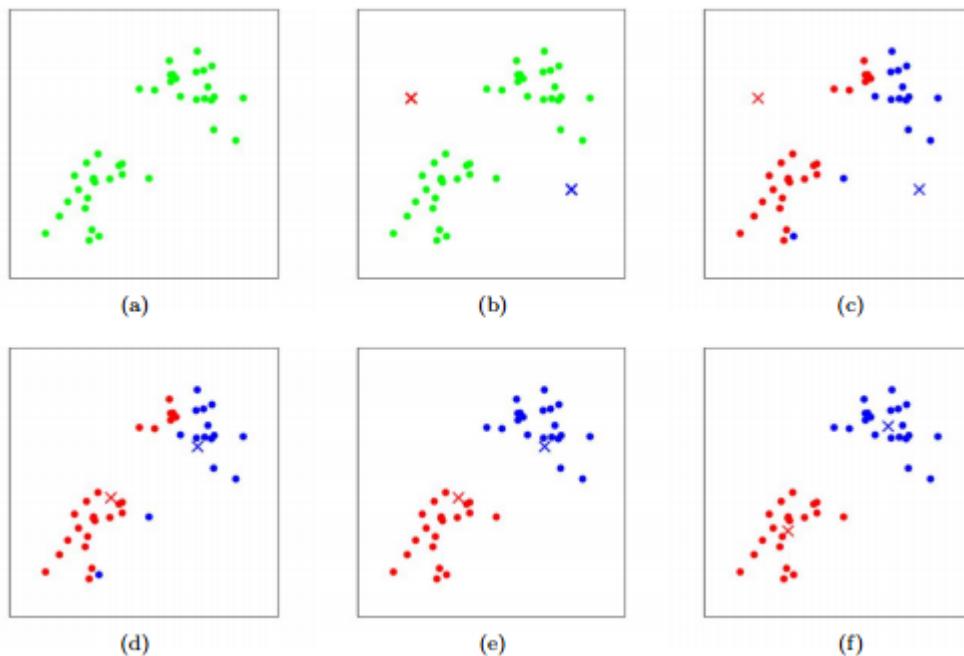


Types of ML

- Unsupervised learning

The data has not been classified previously.

The system should be able to recognize patterns and generate their own labels. Model should classify new input data.

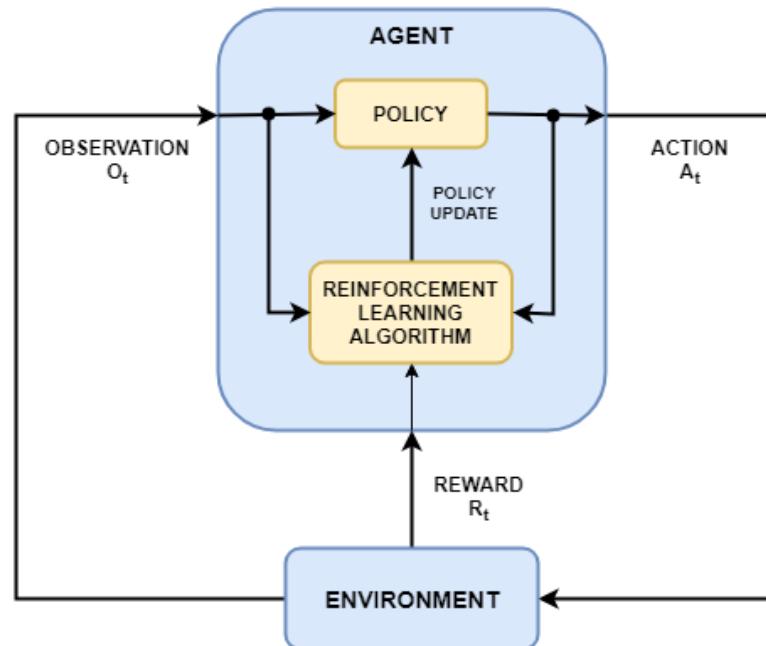


Types of ML

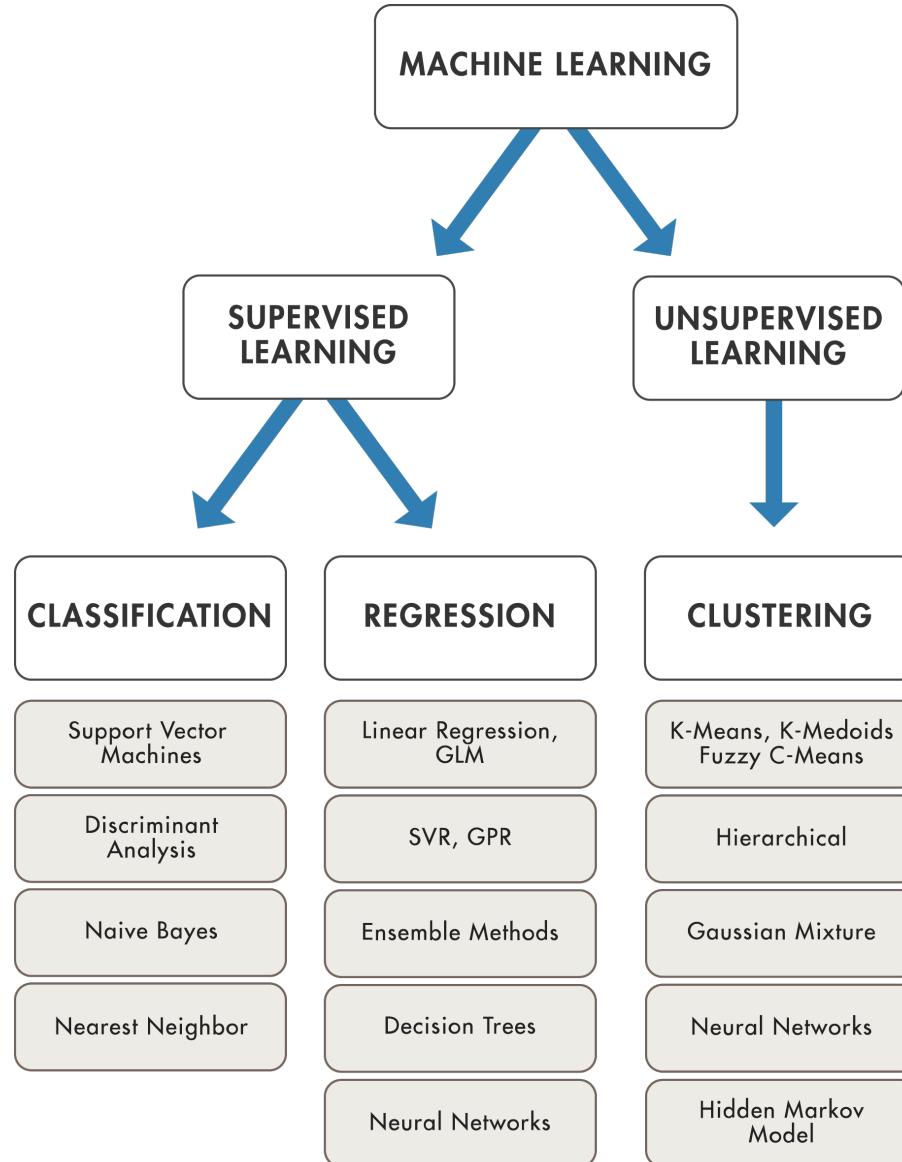
- Reinforcement learning

We have an environment which represents the outside world to the agent and an agent that takes actions.

The agent receives observations from the environment that consists of a reward for his action and information of his new state.



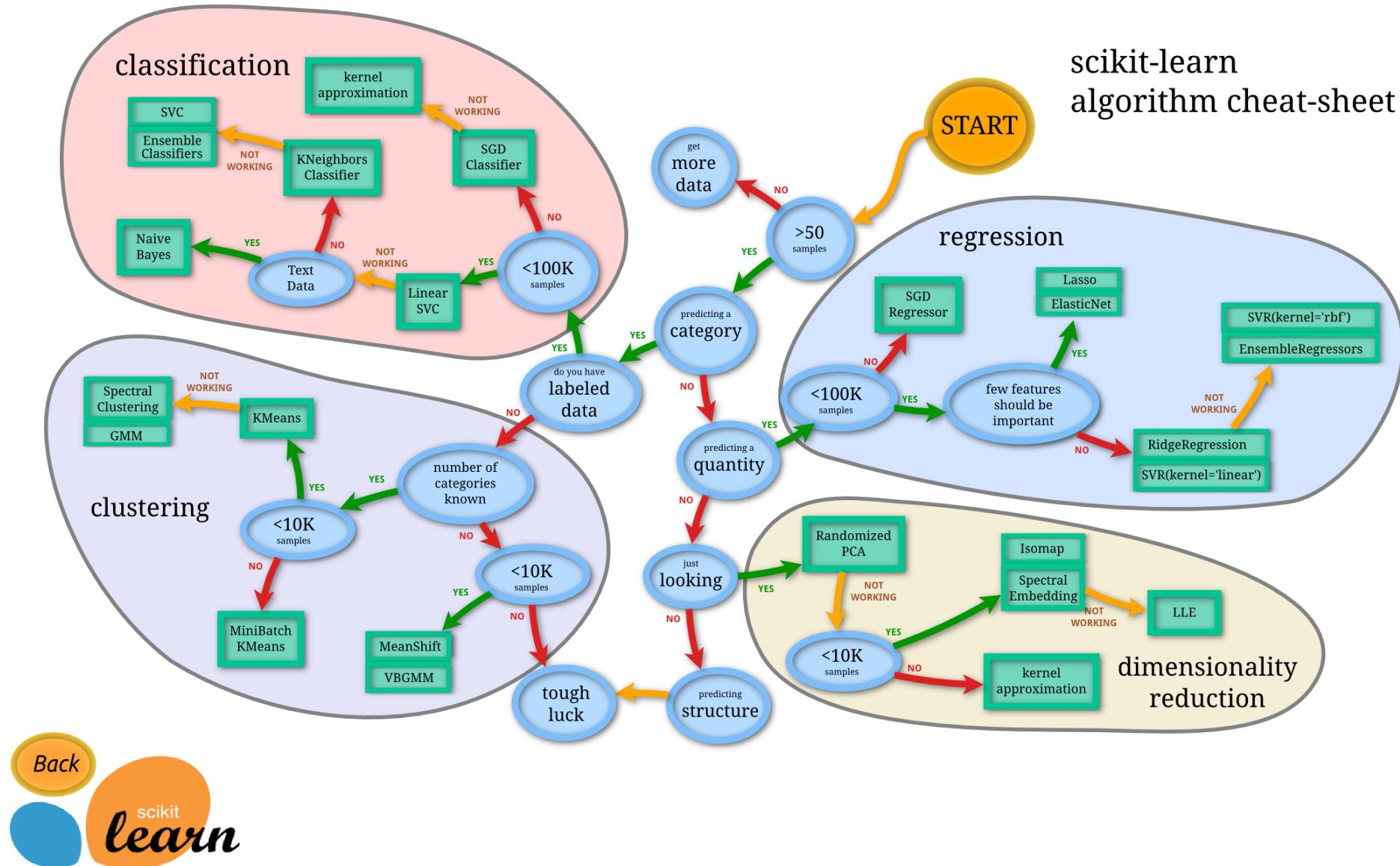
Métodos de machine learning



Librerías en Python

- Pandas
 - <https://pandas.pydata.org/>
 - Lectura de datos, limpieza, ingeniería de features, normalización
- Scikit-learn
 - https://scikit-learn.org/stable/user_guide.html
 - Entrenamiento de modelos supervisados y no supervisados
- Spacy
 - <https://spacy.io>
 - Procesamiento de lenguaje natural

Scikit-learn methods



Computer Vision & Deep Learning

Roberto Muñoz

Senior Data Scientist

roberto.munoz@cl.ey.com

Ricardo Solar

Manager

ricardo.solar@cl.ey.com

Jose Venegas

Senior Manager

jose.venegas@cl.ey.com