

# Azure Machine Learning

Conf.dr. Cristian KEVORCHIAN

Facultatea de Matematică și Informatică

# Azure-platformă pentru cloud computing

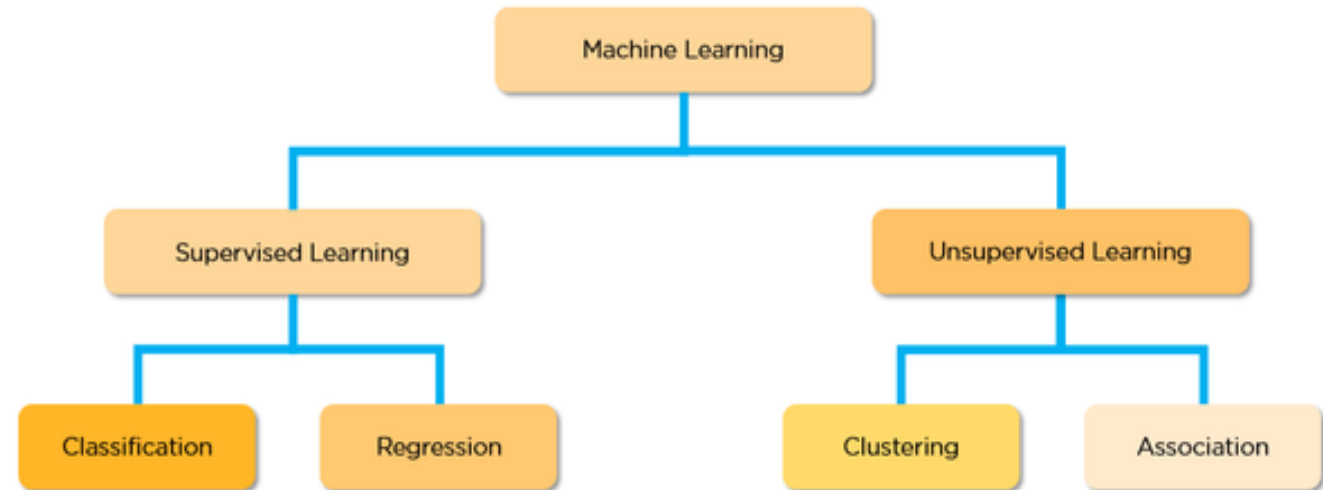
- Dezvoltarea, implementarea și managementul aplicațiilor și serviciilor în regim "cloud computing", Azure are alocată o rețea globală de 54 de regiuni care includ peste 100 de centre de date cu peste 3.7 milioane de servere fizice la nivel global.
- Oferă, platformă ca serviciu (PaaS), infrastructura ca serviciu (IaaS) și software ca serviciu SaaS. Suportă numeroase limbaje de programare, în diverse metodologii de dezvoltare și diverse medii de programare, incluzând atât sisteme software aparținând Microsoft dar și altor vendori (Oracle, IBM, Google, dar și limbaje cum ar fi Python, Java, PHP etc.)
- Utilizează un sistem de operare distribuit în care context se execută "fabric layer", un cluster de dimensiuni mari.
- Scalarea și fiabilitatea sunt asigurate de platforma de calcul distribuit Microsoft Azure Fabric Services, astfel încât serviciile și containerele nu eșuează dacă unul dintre servere este nefuncțional într-unul din centrele de date Microsoft.
- Se adaugă în mod constant noi servicii și funcții - actualizări frecvente la serviciile existente utilizând o abordare de tip Agile

# Machine Learning

- Disciplină științifică care are drept obiectiv de studiu construcția de algoritmi și modele matematice a căror implementare vizează îmbunătățirea progresivă a performanțelor relative la un anumit task. Procesele de învățare automată implementează un model pentru instruire datelor în scopul realizării de predicții sau de a genera decizii fără a fi programat în mod explicit pentru realizarea acestui task. Ex. E-mail filtering, intrusion detection și computer vision, unde este foarte dificil să se dezvolte un algoritm care efectiv să implementeze task-ul.
- Teoria și practica ML este strâns legată de Statistica Computațională, care este un domeniu de frontieră între statistica și computer science.
- Data Mining este un subdomeniu al informaticii. Reprezintă procesul de identificare a tiparelor în volume de date mari care implică utilizarea de euristici la frontiera dintre inteligență artificială, ML, statistică matematică și a sistemelor de baze de date relaționale sau noSQL.

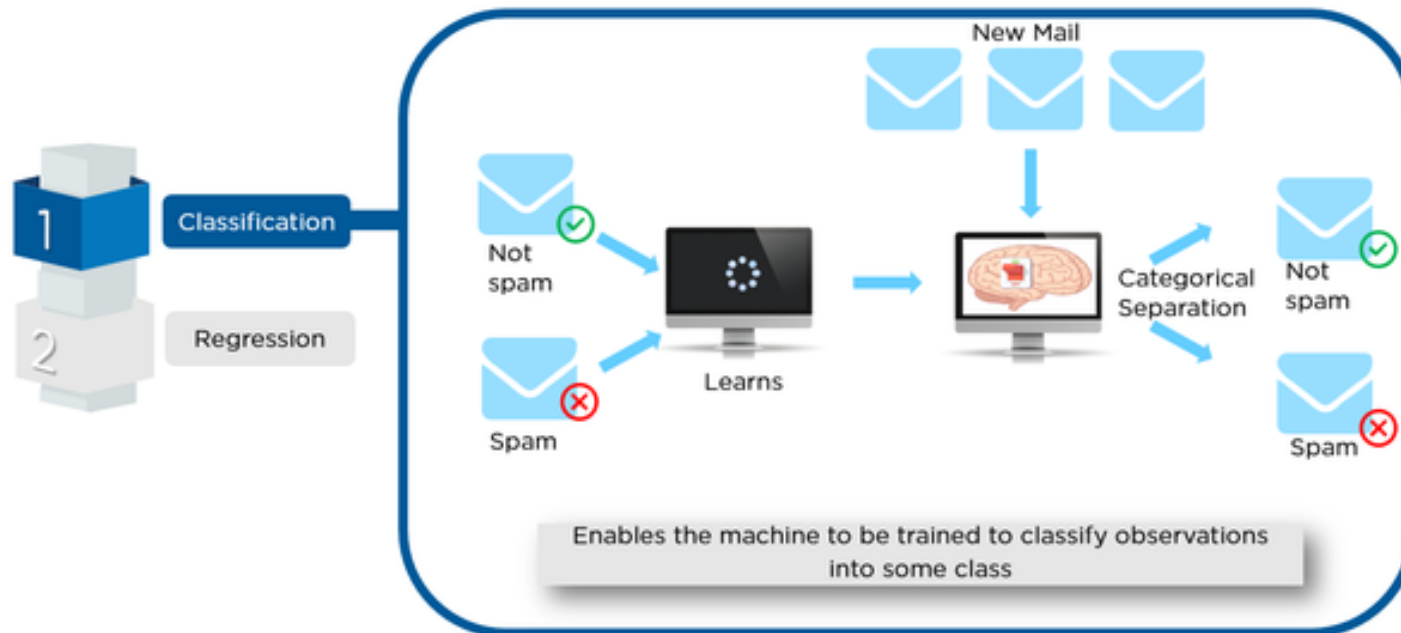
# ML-Tipuri de învățare

- **Învățarea supervizată:** Sistemul este instruit folosind datele anterioare (care includ intrările și ieșirile) fiind recomandat a fi aplicat la probleme de decizie sau pentru realizarea de previziuni atunci când se întâlnesc date noi.
- **Învățare nesupervizată:** Sistemul este capabil să recunoască regularități (tiparele), luând în considerare numai datele de intrare.
- **Învățare consolidată:** deciziile sunt luate de sistem pe baza recompensei / pedepsei aplicate pentru ultima acțiune efectuată.



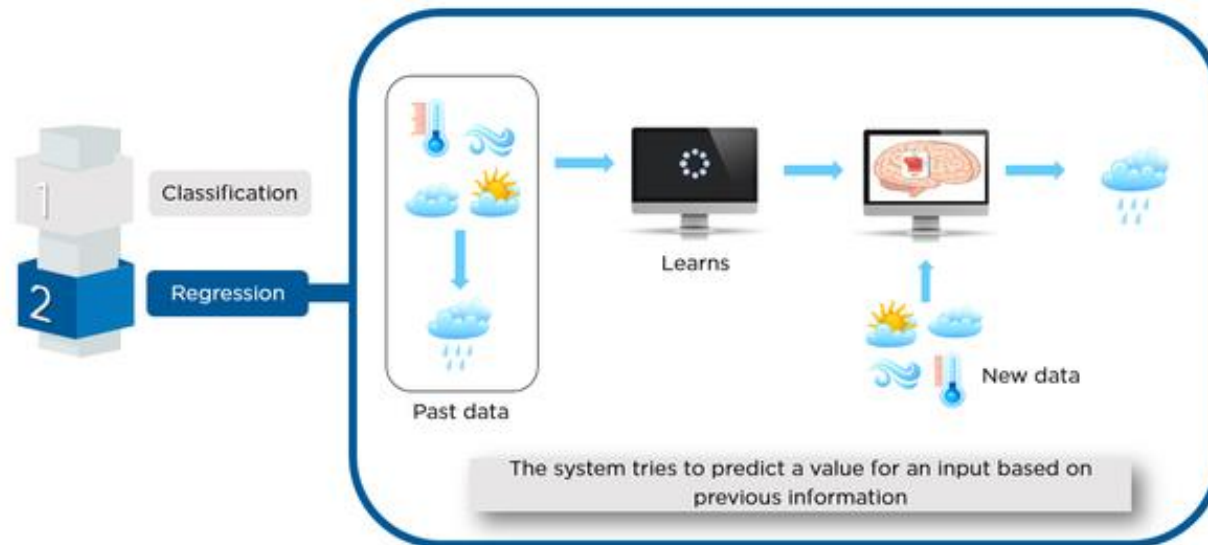
# Clasificarea

- Datele trebuie să fie împărțite într-un anumit număr de categorii bazate pe un proces de "training" folosind date istorice.
- Un algoritm folosit în probleme de clasificare, este teorema naiv Bayes. Clasificatorii naiv Bayes sunt o familie de „clasificatori probabilistici” simpli, bazați pe aplicarea teoremei Bayes cu ipoteze puternice (naive) de independență între caracteristici. Sunt printre cele mai simple modele de rețea bayesiană.



# Regresia

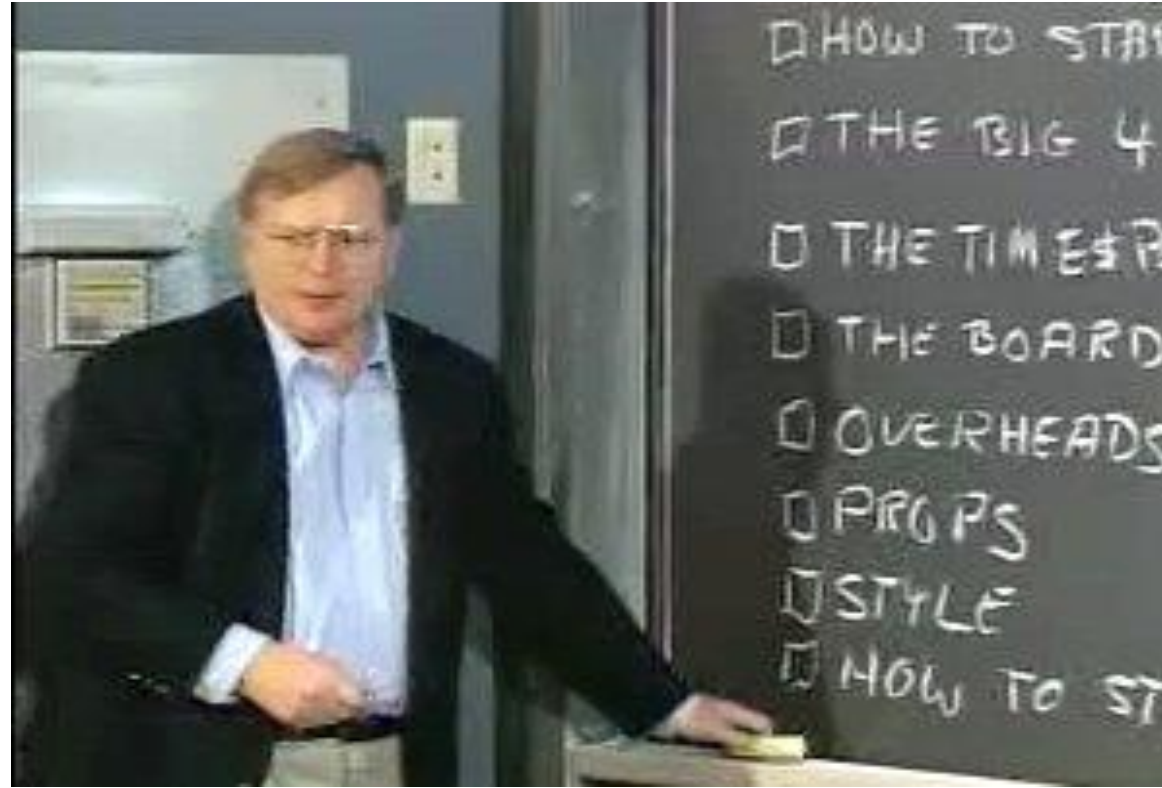
Anticipăm o valoare pentru o intrare pe baza informațiilor primite anterior. Deși acest lucru pare similar clasificării, având în vedere că ambele utilizează date istorice pentru a face previziuni, similitudinea lor se rezumă la acest amănunt. În cazul regresiei, trebuie să estimăm o valoare și nu doar o clasă de observație.



# Încărcarea Modelului de Date și Învățarea

			Variabile			Ce dorești să poți prezice	
			A	B	C		
	Date de Antrenare		1.2	45	14	N	
			1.4	43	16	N	
			1.1	67	3	D	
			1.2	55	7	D	
			1.2	30	13	N	
	Date de test		1.3	43	16	Ar trebui luat N	
			1.1	65	3	Ar trebui luat D	

# Learning, Machines and People → Small Chunks



Patrick Winston, Professor MIT

"You can only learn that which you almost already know."



# Value of ML

- Cu o îmbunătățire de 3% a procentului de detectare a fraudelor cu carduri cadou au fost evitate pierderi de 40 de milioane de dolari. **Nu trebuie ca ceva să fie perfect pentru a fi valoros**
- Viziunea comunității de ML vizează faptul că învățarea automată trebuie democratizată pentru a fi accesibilă oricărei întreprinderi, data scientist, dezvoltator, information worker, utilizator de device-uri oriunde în lume

# Accent pus pe Analitici Predictive

- Punctajul asociat creditului a fost folosit pentru prima dată de companiile de comenzi prin poștă din anii 1950
- Beneficii
  - **Viteza de procesare** – evaluarea a milioane de clienți în secunde
  - **Precizia** - mai exactă decât evaluările efectuate de operatori umani cu aproximativ 20-30%
  - **Consistența** - un model va genera întotdeauna aceeași predicție asociată acelui set de date - evaluarea un expert uman competent va depinde de starea de spirit, sănătate sau mediu

# Exemple de Analitici Predictive

- Identificarea persoanelor care nu își plătesc taxele și impozitele
- Calculul probabilității de a avea un AVC în următorii 10 ani
- Localizarea tranzacțiilor frauduloase cu cărți de credit.
- Selectarea suspectilor în cazuri penale
- Estimarea probabilității ca un client al unei bănci să ajungă în faliment
- Retența clienților
- Probleme de recomandare
- Prognozarea speranței de viață

# Azure Machine Learning (AML)

# Azure ML limitează patru tendințe

- Cheltuieli - costuri enorme de configurare a instrumentelor de lucru, expertiză și capacitate de calcul / stocarea datelor creează obstacole semnificative
- Date Silozate-stocarea localizată și gestionarea greoaie a datelor inhibă accesul la date
- Instrumente deconectate
- Instrumentele complexe limitează participarea la explorarea datelor și a modelelor de construire
- Complexitate de implementare
- Multe modele nu oferă niciodată valoare adăugată de business din cauza problemelor de implementare în producție

# Microsoft Azure Machine Learning Studio

Medii de dezvoltare colaborativă care permit efectuarea de:

- Build
- Test
- Deploy

AI/ML și DEVOPS operează împreună cu success.

# Azure ML Studio

Microsoft Azure Machine Learning

Home

Studio

Gallery PREVIEW



EXPERIMENTS



WEB SERVICES



DATASETS



TRAINED MODELS



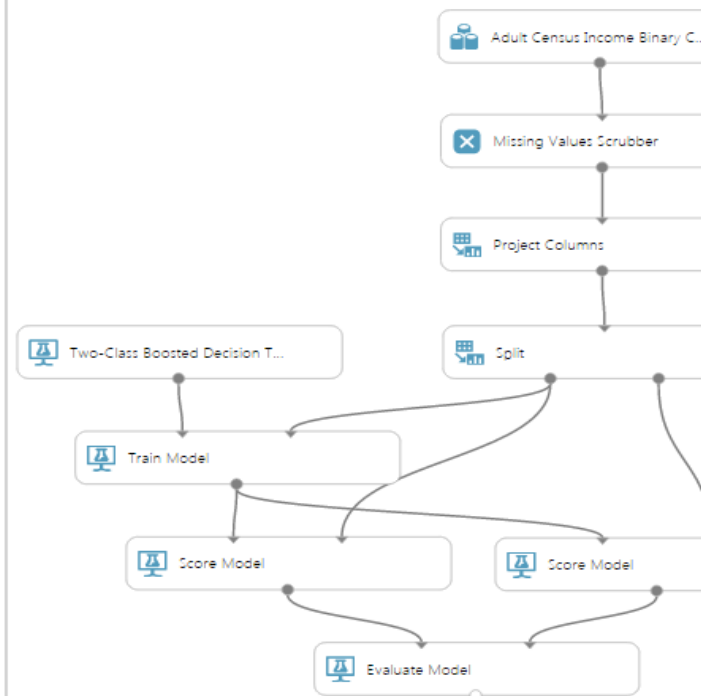
SETTINGS

## experiments

MY EXPERIMENTS

SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED
<input checked="" type="checkbox"/>	Sample 5: ScoreExperiment	alannoel	Draft	2/7/2015 6:40:25 PM
<input type="checkbox"/>	Chapter 02 - Hello ML AI Noel	alannoel	Finished	12/3/2014 2:44:05 PM
<input type="checkbox"/>	Chapter 02 - Hello ML AI Noel	alannoel	Draft	12/3/2014 2:32:14 PM
<input type="checkbox"/>	Sample Network Intrusion Detect...	alannoel	Finished	9/6/2014 10:37:59 AM
<input type="checkbox"/>	Sample Experiment - Demand Fo...	alannoel	Draft	7/25/2014 4:42:53 PM
<input type="checkbox"/>	German Credit Prediction	alannoel	Draft	7/25/2014 4:18:41 PM
<input type="checkbox"/>	Sample 5: Train, Change to Boost...	alannoel	Finished	7/24/2014 6:53:54 PM
<input type="checkbox"/>	Japanese Trained Model	alannoel	Draft	7/24/2014 6:51:53 PM
<input type="checkbox"/>	Untitled	alannoel	Draft	7/24/2014 6:33:09 PM
<input type="checkbox"/>	Untitled	alannoel	Finished	7/24/2014 6:13:58 PM
<input type="checkbox"/>	Untitled	alannoel	Finished	7/24/2014 6:07:17 PM
<input type="checkbox"/>	Untitled	alannoel	Draft	7/24/2014 6:02:34 PM
<input type="checkbox"/>	Sample 5: TrainedModel	alannoel	Draft	7/24/2014 4:23:23 PM
<input type="checkbox"/>	Sample 5: TrainCopy	alannoel	Finished	7/24/2014 4:21:47 PM
<input type="checkbox"/>	Income Level Prediction	alannoel	Finished	7/24/2014 1:12:12 PM
<input type="checkbox"/>	Flight Delay Prediction	alannoel	Draft	7/23/2014 8:32:21 PM
<input type="checkbox"/>	Untitled	alannoel	Finished	7/23/2014 6:25:35 PM
<input type="checkbox"/>	Writer01	alannoel	Failed	7/23/2014 6:22:41 PM
<input type="checkbox"/>	Untitled	alannoel	Draft	7/23/2014 6:05:33 PM
<input type="checkbox"/>	Demo01	alannoel	Draft	7/23/2014 6:00:41 PM



# Portalul pentru Management

Microsoft Azure | alannoel@msn.com

**STORAGE** 5

**HDINSIGHT** 0

**MEDIA SERVICES** 0

**SERVICE BUS** 0

**VISUAL STUDIO ONLINE** 1

**CACHE** 0

**BIZTALK SERVICES** 0

**RECOVERY SERVICES** 0

**CDN** 0

**AUTOMATION** 0

**SCHEDULER** 0

**API MANAGEMENT** 0

**MACHINE LEARNING** 1

**NETWORKS** 1

**TRAFFIC MANAGER** 0

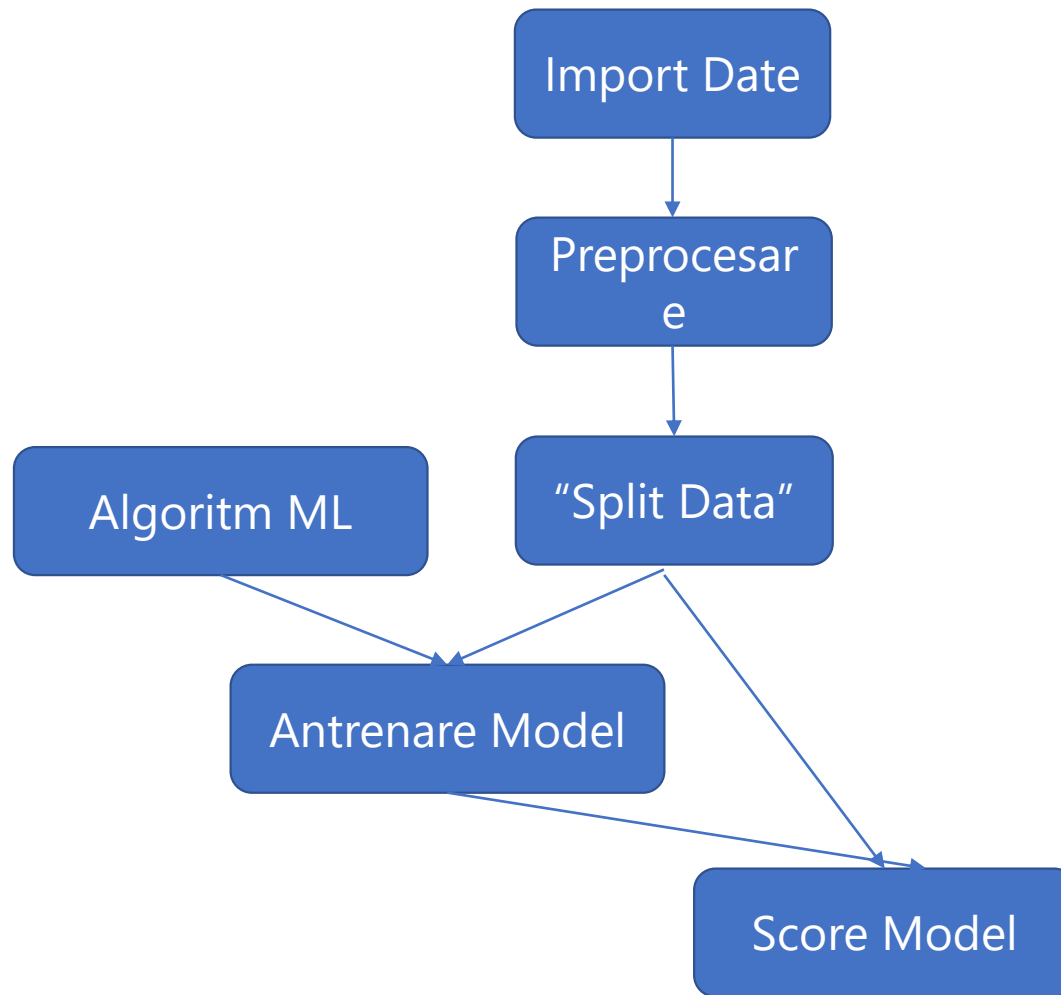
**REMOTEAPP**

all items

NAME	TYPE	STATUS	SUBSCRIPTION	LOCATION	
alnoel01	Storage Account	✓ Online	Pay-As-You-Go	East US	
alnoelws01	Storage Account	✓ Online	Pay-As-You-Go	South Central US	
anoelw2k12	Storage Account	✓ Online	Pay-As-You-Go	East US	
portalvhds4vxb6nz1fsc	Storage Account	✓ Online	Pay-As-You-Go	North Central US	
portalvhds20ws6qp6033k	Storage Account	✓ Online	Pay-As-You-Go	East US	
alnoel01	Cloud service	✓ Running	Pay-As-You-Go	North Central US	
anoelW2K12	Cloud service	✓ Running	Pay-As-You-Go	East US	
anoelW2K12	Virtual machine	✓ Running	Pay-As-You-Go	East US	
wolfie	Website	■ Stopped	Pay-As-You-Go	East US	
alnoel12feb	Website	■ Stopped	Pay-As-You-Go	East US	
Group Group anoelW2K12	Virtual Network	✓ Created	Pay-As-You-Go	East US	
Default Directory	Directory	✓ Active	Shared by all Default Directory subscriptions	United States	
alnoel	Visual Studio Online	✓ Active	Pay-As-You-Go	North Central US	
alnoelws01	ML Workspace	✓ Online	Pay-As-You-Go	-	



# Fluxul de procesare în ML



# Probleme în obținerea unor predicții de calitate

- Cantități de date adecvate— nu este neapărat necesar să avem volume mari de date.
- Etichete
  - Exemple de etichetare
    - Reteția clienților: înregistrarea clienților loiali + înregistrarea clienților ce părăsesc firma(churn)
    - Trăsături relevante
      - Informații despre clienți: vârstă, sex, cod poștal, tipare de cheltuieli anterioare
      - Informații despre tranzacție: suma, tranzacțiile anterioare
    - Toleranță la incertitudine
  - Nu trebuie ținută perfecțiunea(îmbunătățirea detectării fraudei cu 3% = milioane de dolari)

# Variety of Applications of Azure ML



## Machine Learning in ML Studio

## Anomaly Detection

One-class Support Vector Machine  
Principal Component Analysis-based Anomaly Detection  
Time Series Anomaly Detection\*

## Classification

## Two-class Classification

Averaged Perceptron  
Bayes Point Machine  
Boosted Decision Tree  
Decision Forest  
Decision Jungle  
Logistic Regression  
Neural Network  
Support Vector Machine

## Multi-class Classification

Decision Forest  
Decision Jungle  
Logistic Regression  
Neural Network  
One-vs-all

## Clustering

K-means Clustering

## Recommendation

Matchbox Recommender

## Regression

Bayesian Linear Regression  
Boosted Decision Tree  
Decision Forest  
Fast Forest Quantile Regression  
Linear Regression  
Neural Network Regression  
Ordinal Regression  
Poisson Regression

## Statistical Functions

Descriptive Statistics  
Hypothesis Testing T-Test  
Linear Correlation  
Probability Function Evaluation

## Text Analytics

Feature Hashing  
Named Entity Recognition  
Vowpal Wabbit

## Computer Vision

OpenCV Library

## Data/Model Visualization

- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree\*

## Training

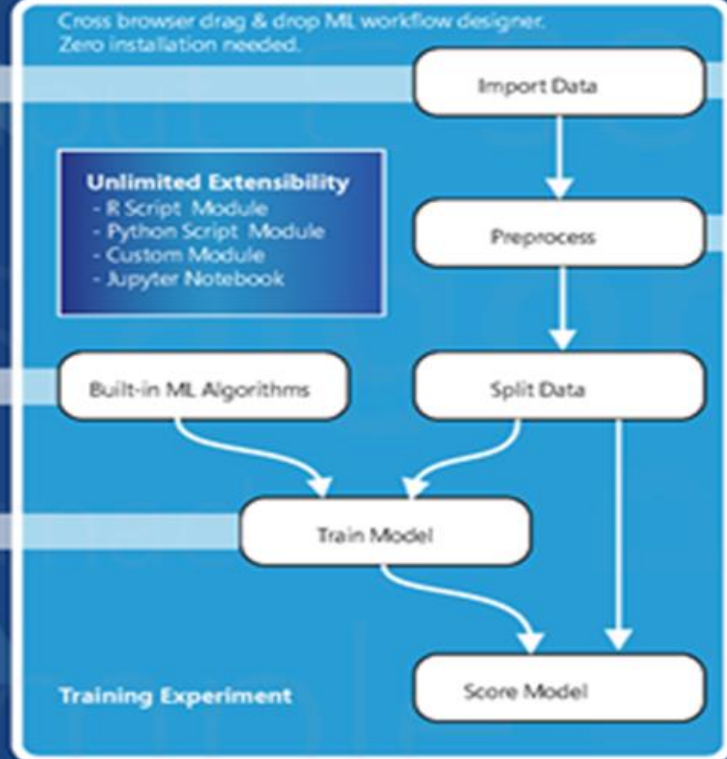
- Cross Validation
- Retraining
- Parameter Sweep

<https://studio.azureml.net>

Guest Access Workspace: Free trial access without logging in.

Free Workspace: Free persisted access, no Azure subscription needed.

Standard Workspace: Full access with SLA under an Azure subscription.



## Data Source

- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW\*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server\*
- Web URL (HTTP)

## Data Format

- ARFF
- CSV
- SVMlight
- TSV
- Excel
- ZIP

## Data Preparation

- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

## Enterprise Grade Cloud Service

- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance\*

## Community

- Gallery (<http://gallery.azureml.net>)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

## One-click Operationalization

Predictive Experiment

## Make Prediction with Elastic APIs









- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

\* Feature Coming Soon

## Azure Machine Learning Studio Capabilities Overview

DEMO



 NEW
  RUN HISTORY
  SAVE
  SAVE AS
  DISCARD CHANGES
  RUN
  SET UP WEB SERVICE
  PUBLISH TO GALLERY