ETF Working together Learning for life

Session 5

Exploring the knowledge and lessons from ETF project Big Data for LMI

Compact presentation of the technical construction of the data system. Focus on data collection, data classification and visualisation

Mauro Pelucchi

22-24 November 2021

Emsi | burning glass

Topics

1. What is Machine Learning?
2. Databricks (intro)
3. Design a pipelines
    1. How to scrape online job vacanciesBuild our pipeline with Spark
    2. Focus on occupation’s categorization

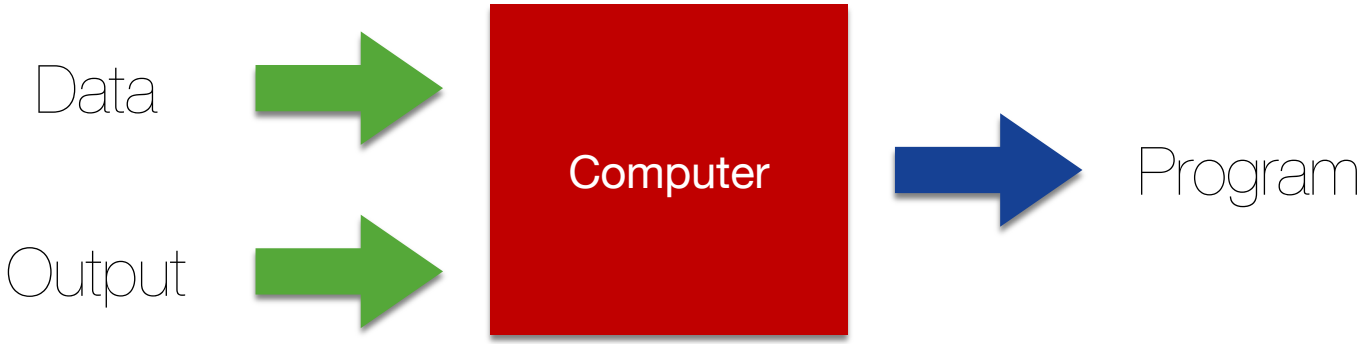
Topics

1. **What is Machine Learning?**
2. Databricks (intro)
3. Design a pipelines
    1. How to scrape online job vacanciesBuild our pipeline with Spark
    2. Focus on occupation’s categorization

Machine Learning

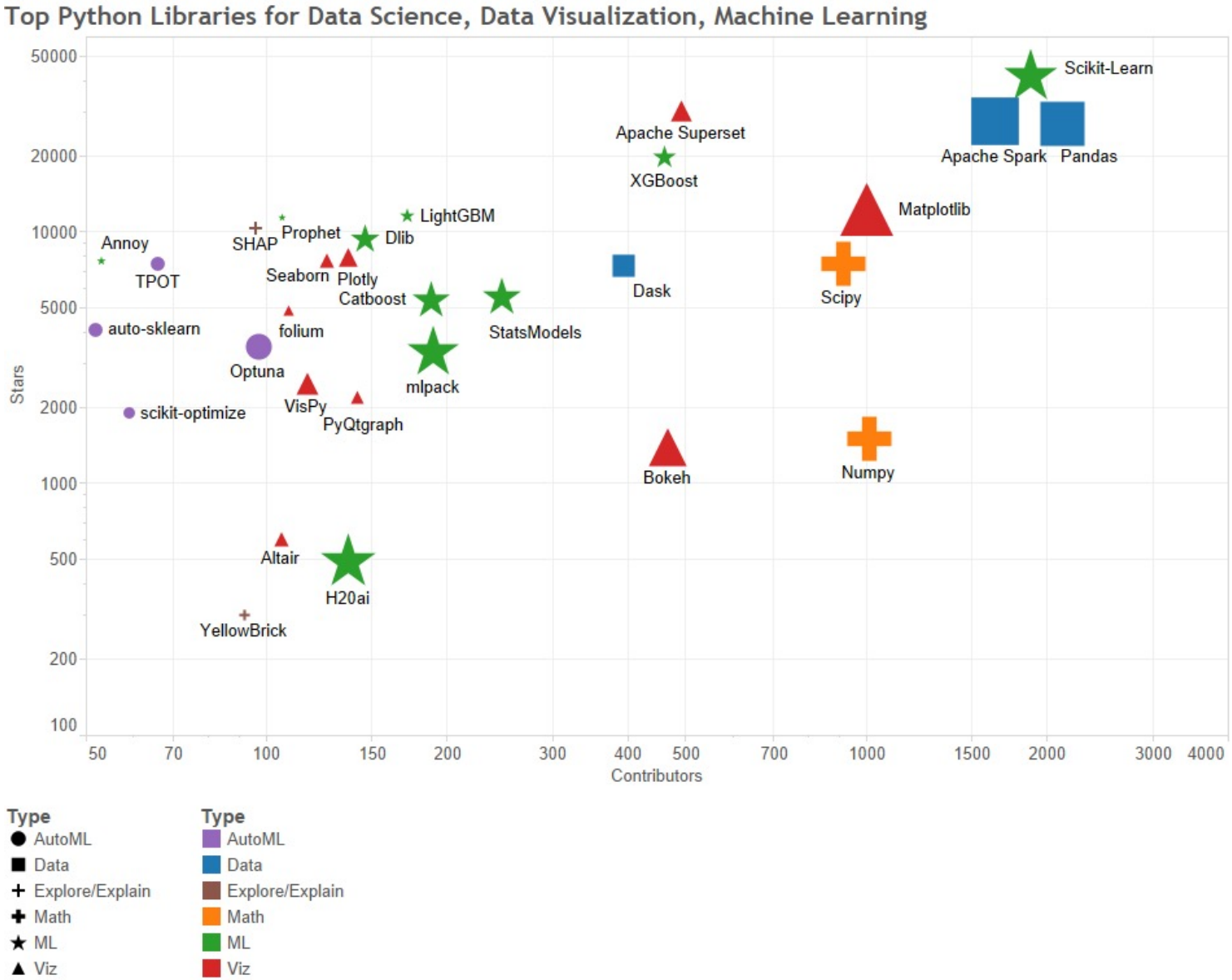
Learning is any process by which a system improves performance from experience.

Herbert Simon



Machine Learning Definition

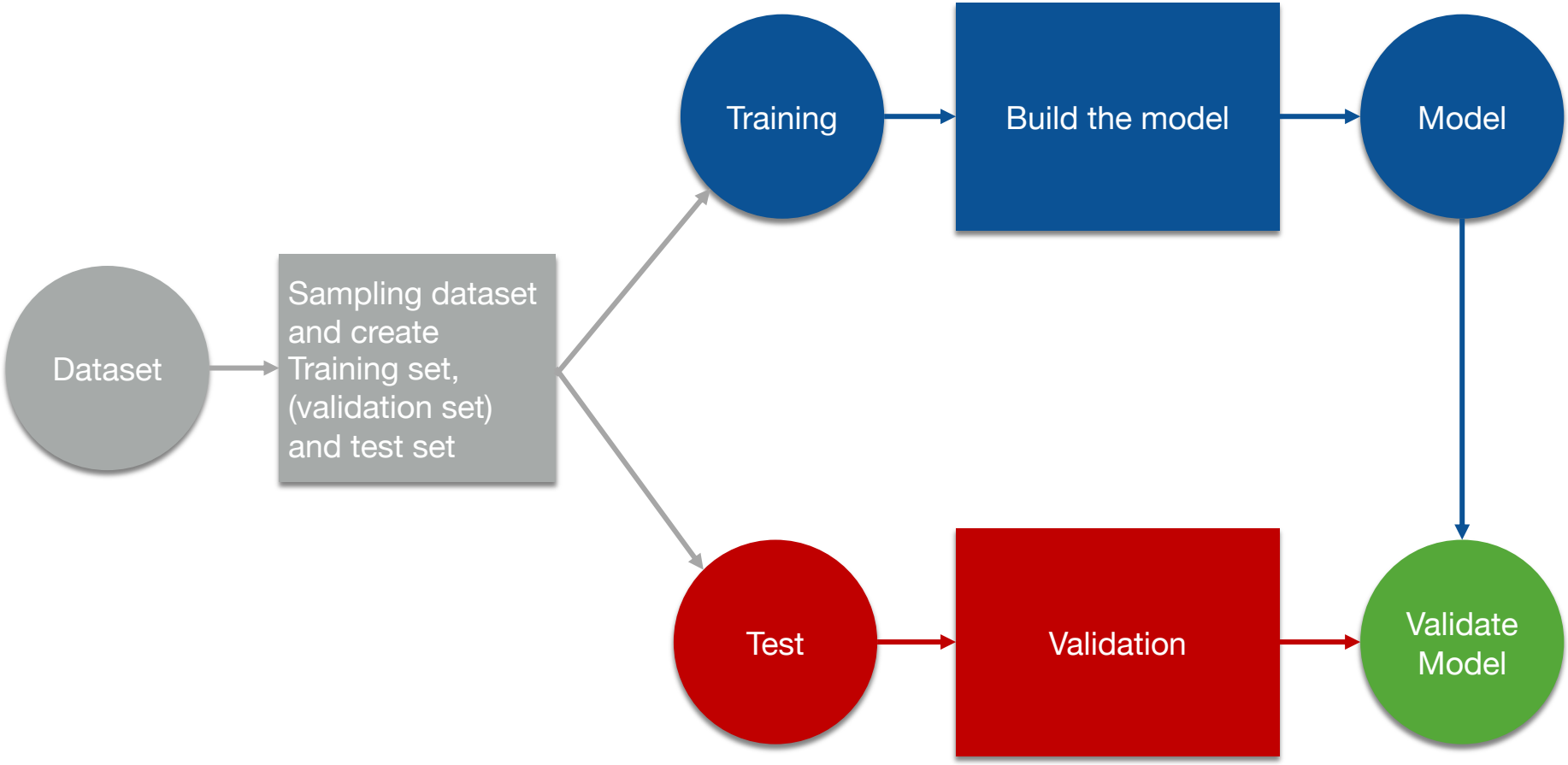
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.



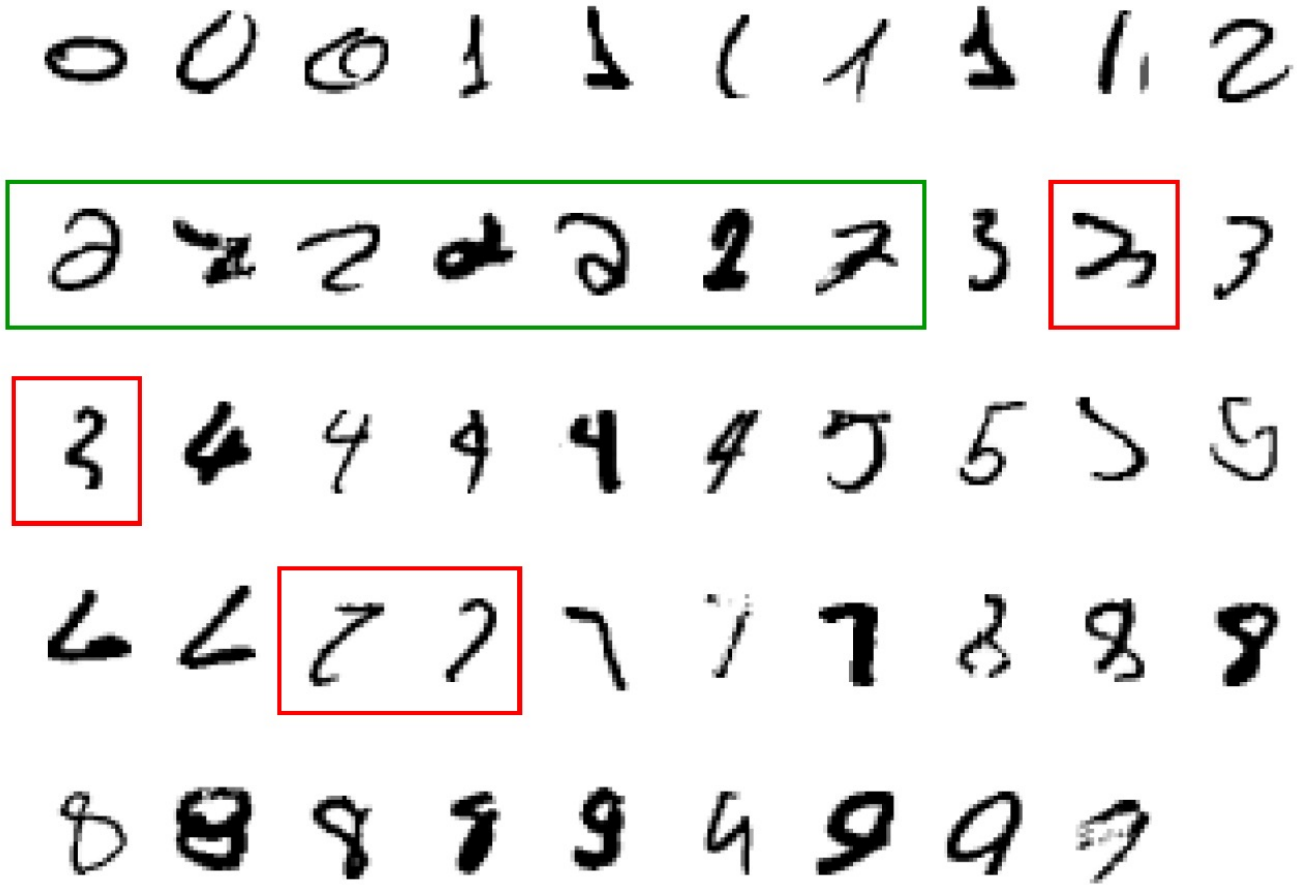
https://www.kdnuggets.com/2020/11/top-python-libraries-data-science-data-visualization-machine-learning.html

Machine learning

Training is the process of making the system able to learn.



A classic example of a task that requires machine learning:
It is very hard to say what makes a 2



Slide credit: Geoffrey Hinton

Type of learning

Supervised (inductive) learning

- Given: training data + desired outputs (labels)

Unsupervised learning

- Given: training data (without desired outputs)

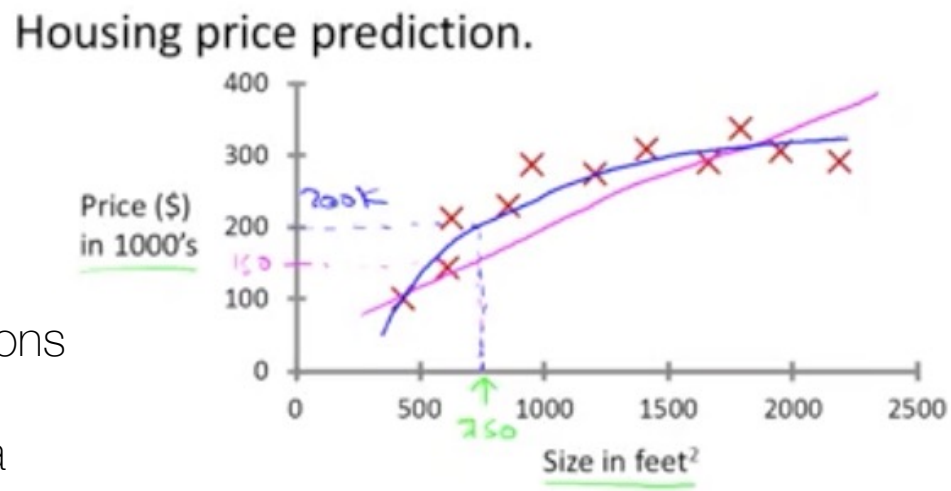
Semi-supervised learning

- Given: training data + a few desired outputs

Reinforcement learning

- Rewards from sequence of actions

Supervised Learning



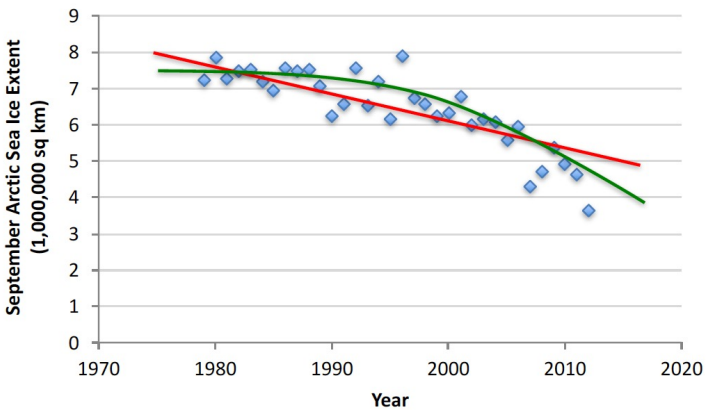
We know the class to which the observations belong

Classification problem: what class does a new observation belong to?

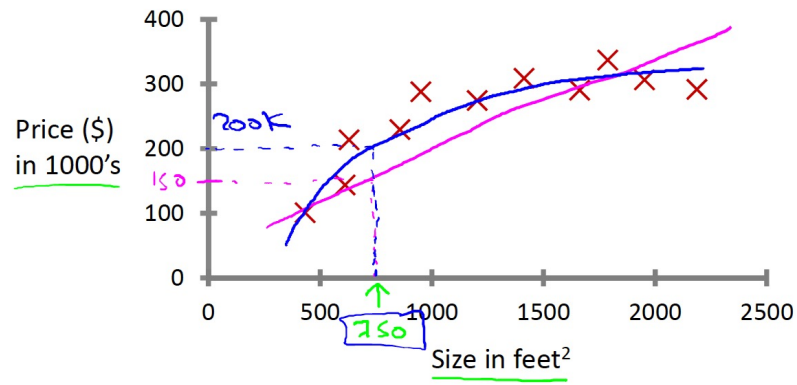
Size in feet^2	Number of rooms	Year of construction	Price ($)
500	3	1983	100.000
1000	4	2005	165.000
1000	3	2016	230.000

Supervised Learning
Regression/Prediction

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Learn a function f(x) to predict y given x
y is **real-valued** —> **regression/prediction**

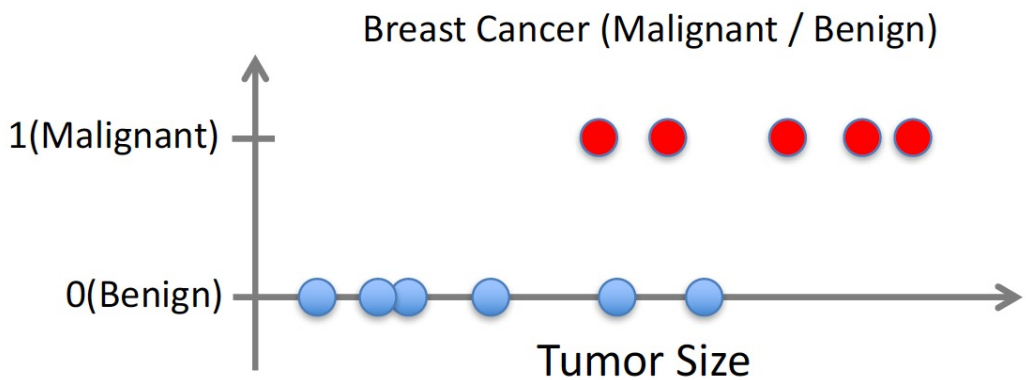


Housing price prediction.

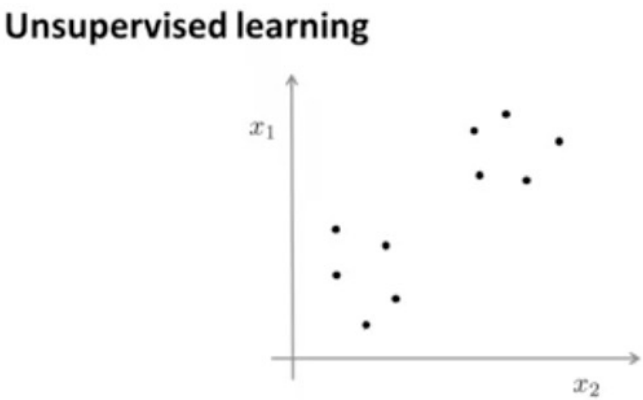


Supervised Learning
Classification

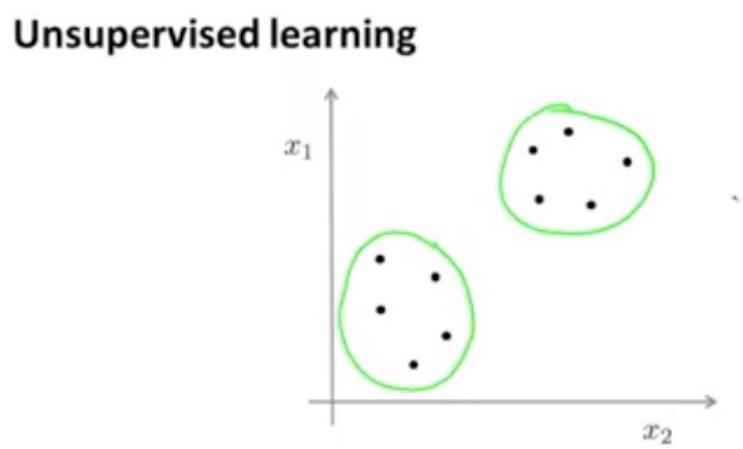
Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Learn a function f(x) to predict y given x
y is **categorical** —> **classification**

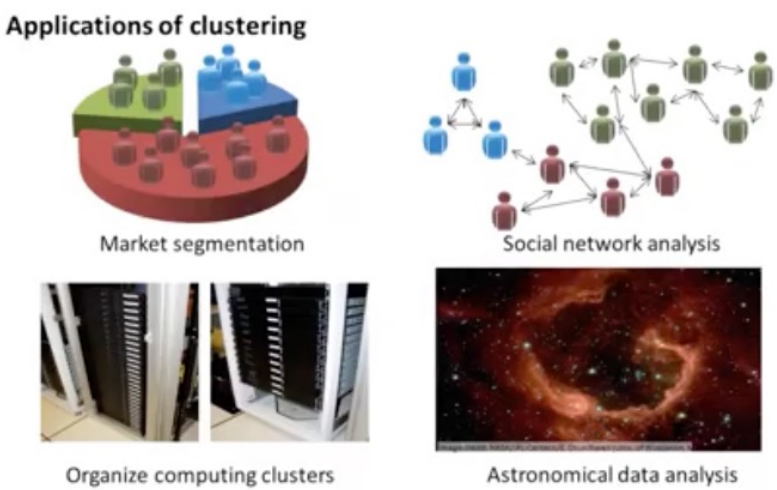


Unsupervised Learning



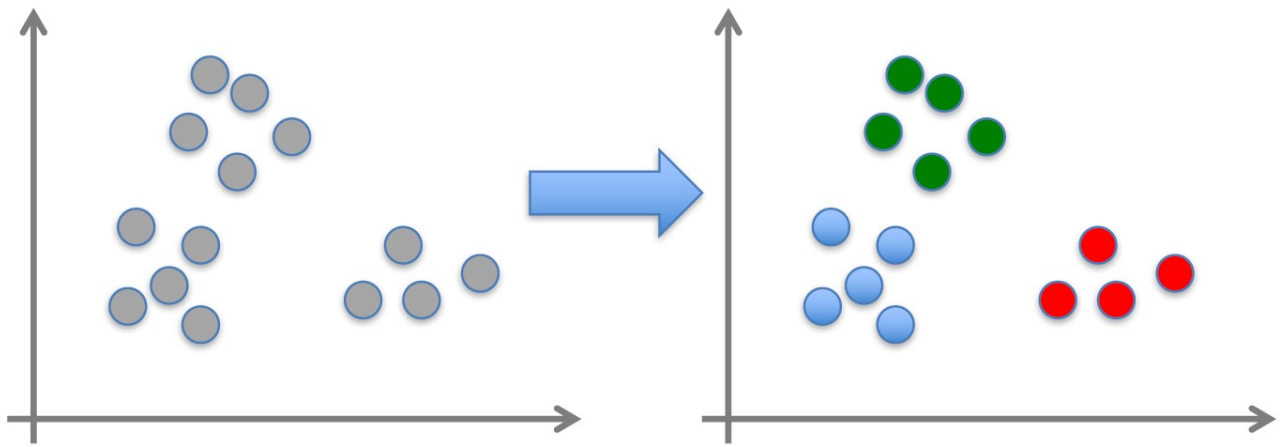
We have no information about the class to which my observations belong.
We look for new features hidden in our data and try to interpret them.





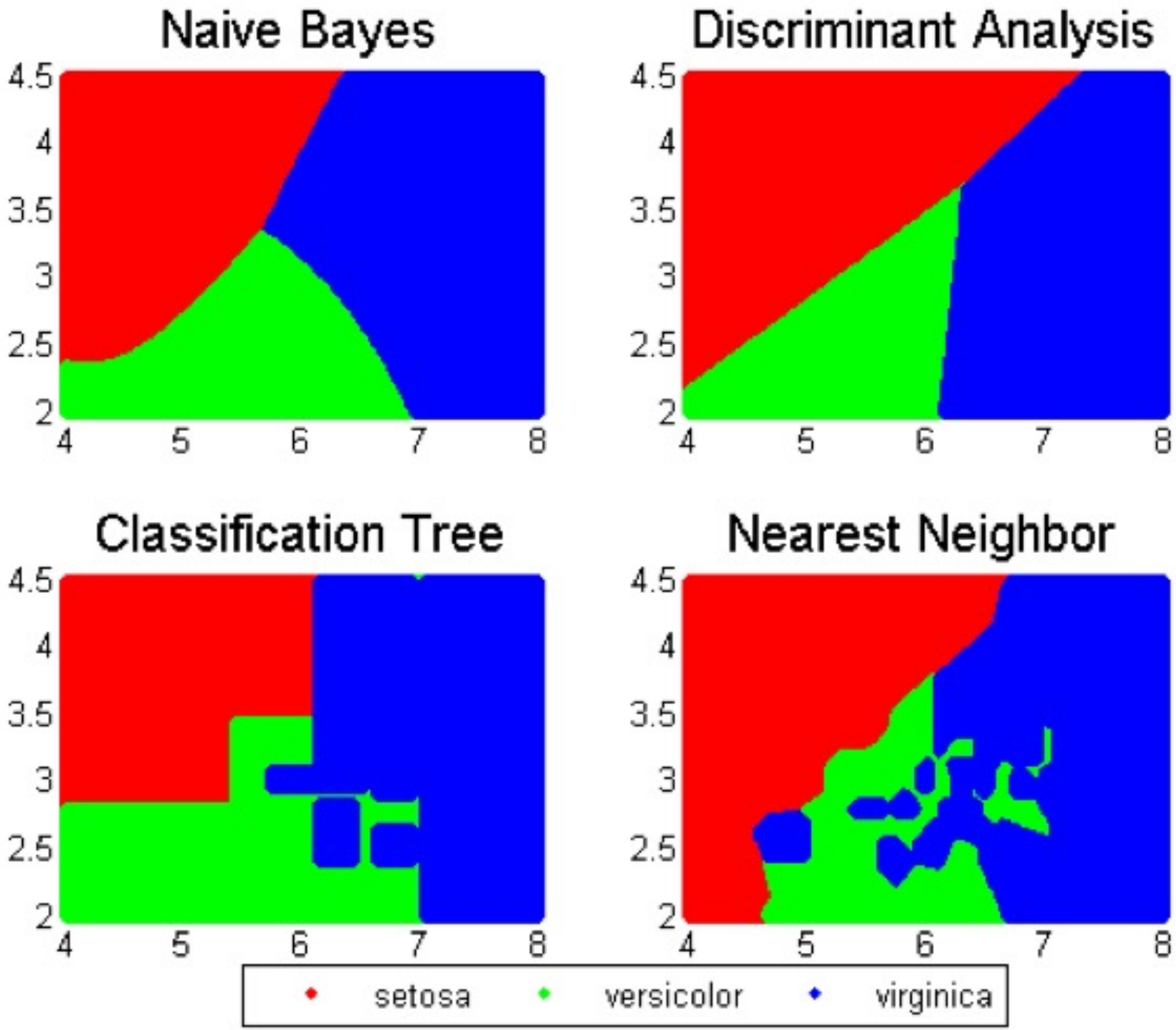
Unsupervised Learning
Classification

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (without labels)
Output hidden structure behind the x's
E.g. —> **clustering, probability distribution estimation, finding association (in features), dimension reduction**



Four different steps

1. Building the regression or classification **models** from a sample or data set, for which the values of both the explanatory variables (or features) Xi and the dependent variable Y are observed/known
2. **Assessment of the performances** of the different models on an independent data set: a **validation set** that was not used for building the models
3. **Evaluation of the performances of the best model**, on an independent data set
4. **Application of the best model to new cases (Score)**


Naive Bayes
Discriminant Analysis
Classification Tree
Nearest Neighbor
4.5
4
3.5
3
2.5
2
4
5
6
7
8
setosa
versicolor
virginica

Performance measures

Predicted class

Actual class

	P	**N**
P	True Positives (TP)	False Negatives (FN)
N	False Positives (FP)	True Negatives (TN)

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

how many of the returned documents are correct (> 0.6)

$$\text{Recall} = \frac{tp}{tp + fn}$$

how many of the positives does the model return (> 0.6)

F-measure

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Performance metric
Precision

It's important **to get error results as a single, numerical value**. Otherwise it is difficult to assess your algorithm's performance.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Precision: how many of the classified documents are correct

(high precision = no garbage)

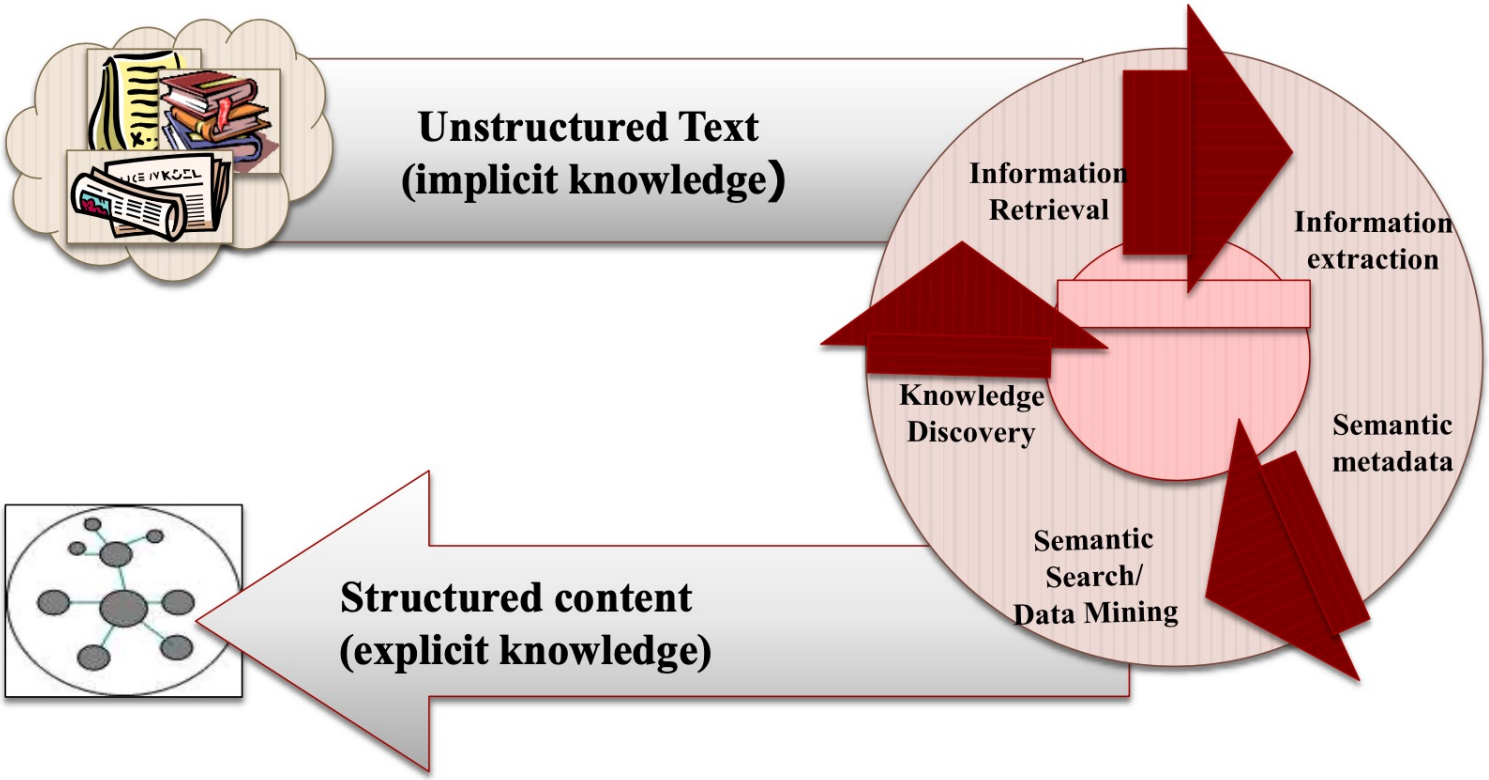
Of all cases we predicted where y=software developer, **what fraction actually is a software developer**?

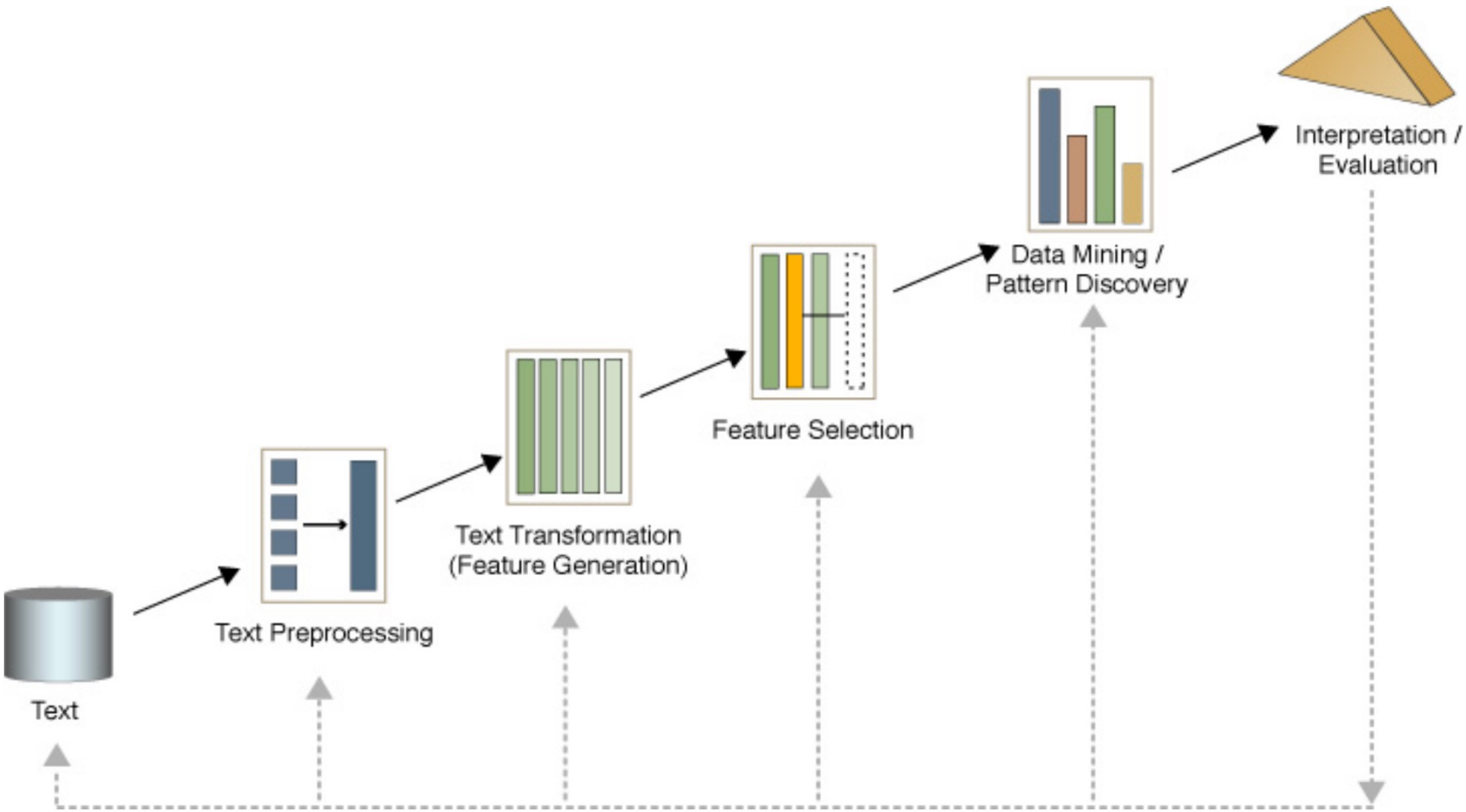
Definition

- Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text.

- Text Mining can be defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools.

- Text Mining seeks to extract useful information from data sources (document collections) through the identification and exploration of interesting patterns.


Unstructured Text
(implicit knowledge)
Information
Retrieval
Information
extraction
Knowledge
Discovery
Semantic
metadata
Semantic
Search/
Data Mining
Structured content
(explicit knowledge)


Text
Text Preprocessing
Text Transformation
(Feature Generation)
Feature Selection
Data Mining /
Pattern Discovery
Interpretation /
Evaluation

Structuring Textual Information

Many methods designed to analyze structured data

If we can represent documents by a set of attributes we will be able to use existing data mining methods

Use statistics to add a numerical dimension to unstructured text

How to represent a document?

- Vector based representation —> Bag of words
- Term frequency (TF)
- Document frequency (DF)
- TF-IDF
- Document length

Weighting Scheme for Term Frequencies

TF-IDF

$$\text{TF-IDF}(w, d) = \text{TermFreq}(w, d) \cdot \log (N / \text{DocFreq}(w))$$

TermFreq(w, d): frequency of w in the document d
N: number of documents in the collection
DocFreq(w): number of documents in the collection that contains w

Weighting Scheme for Term Frequencies

TF-IDF

$$\text{TF-IDF}(w, d) = \text{TermFreq}(w, d) \cdot \log (N / \text{DocFreq}(w))$$

TermFreq(w, d): frequency of w in the document d
N: number of documents in the collection
DocFreq(w): number of documents in the collection that contains w

A term that appears many times in a document receives a high TF-IDF value
if it is not common within the entire document collection:
are RARE and IMPORTANT terms

Weighting Scheme for Term Frequencies

TF-IDF

$$\text{TF-IDF}(w, d) = \text{TermFreq}(w, d) \cdot \log (N / \text{DocFreq}(w))$$

TermFreq(w, d): frequency of w in the document d
N: number of documents in the collection
DocFreq(w): number of documents in the collection that contains w

Terms with low TF-IDF are either infrequent terms in the documents or very common in the collection.
common in the collection: COMMON WORDS AND NOISE

Topics

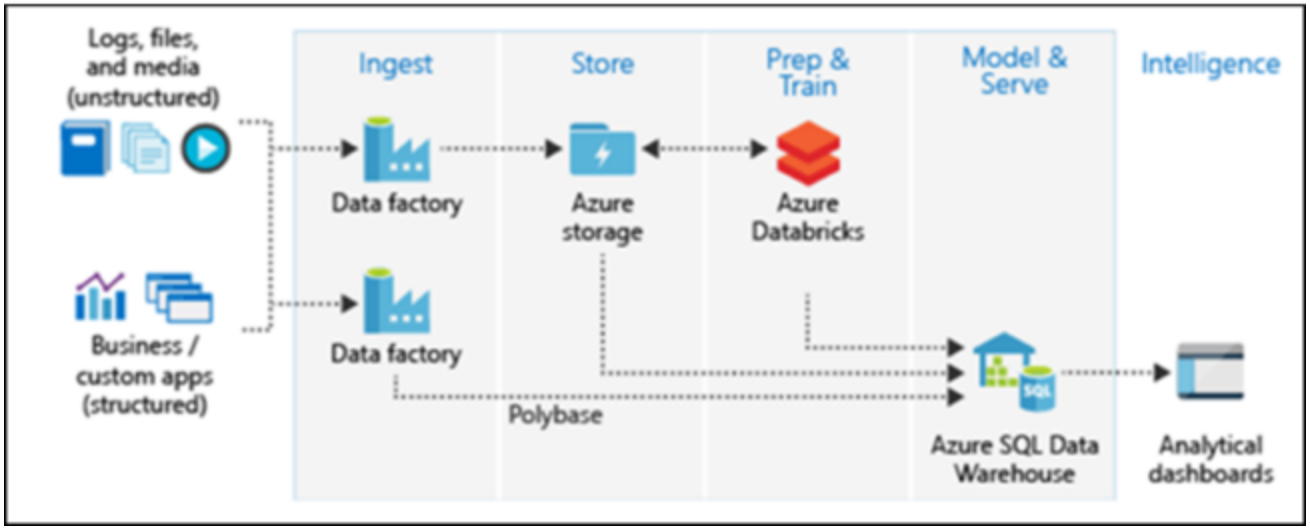
1. What is Machine Learning?
2. **Databricks (intro)**
3. Design a pipelines
    1. How to scrape online job vacanciesBuild our pipeline with Spark
    2. Focus on occupation’s categorization

Databricks

- It's an easy and collaborative analytics platform based on Apache Spark
- All Spark modules are present in Databricks (SparkSQL, Streaming, ML, GraphX)
- "Main goal: to remove all the hardness and complexity to get and manage a Spark cluster”
- Enables one-click installation and settings management
- Offers simplified workflows and interactive workspace to facilitate collaboration between data scientists, developers and business analysts
- Integration with leading cloud providers such as Amazon AWS and Microsoft Azure

Databricks


User Views / Clicks
Impressions
Conversions
User Conversions
Amazon Kinesis / Kafka
databricks
ETL, Aggregation, Attribution
Real-Time Reporting in BI Tools


Logs, files, and media (unstructured)
Business / custom apps (structured)
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Data factory
Azure storage
Azure Databricks
Data factory
Polybase
Azure SQL Data Warehouse
Analytical dashboards

Databricks

Clusters

Notebooks

Jobs

Data

Databricks Notebook

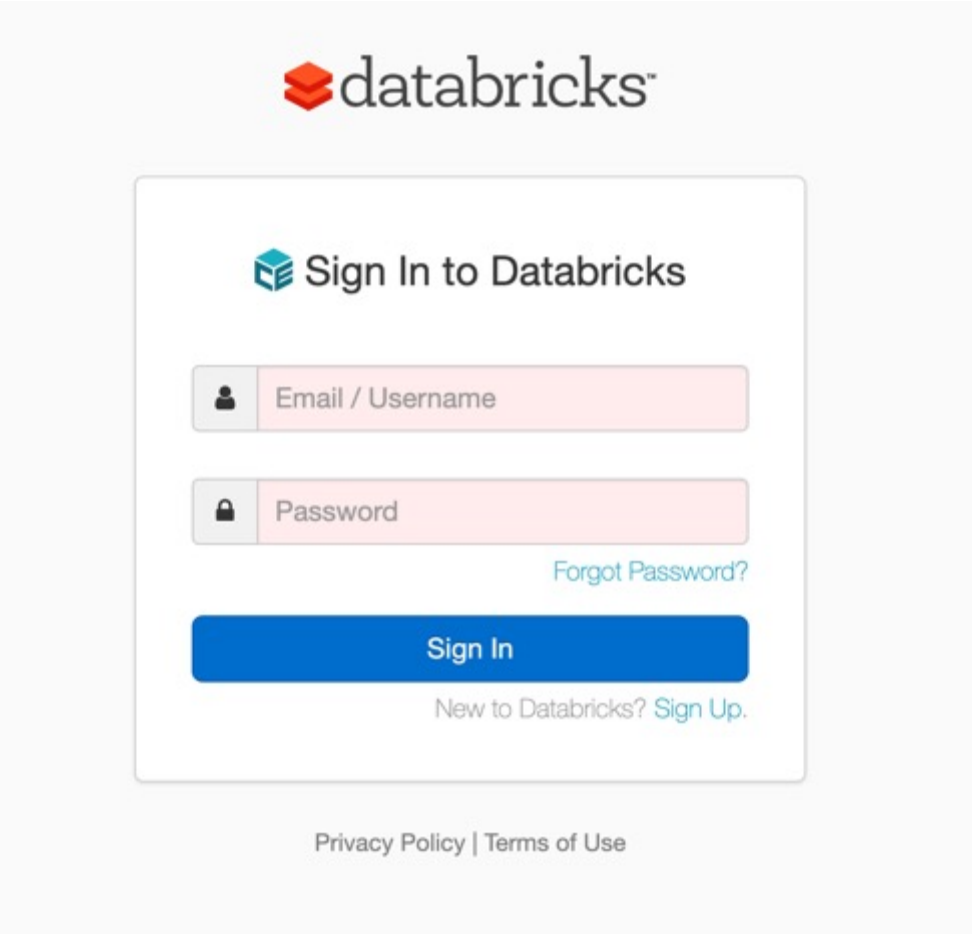
- Similar to Jupyter or Zeppelin notebooks
- Supported languages
  - Python, Scala and SQL (also R…)
  - They can all be used in a single notebook.
- The Spark session is already defined for each notebook as a global spark variable.
- Once a notebook is created it must be connected to an active cluster.

Databricks
Version and collaboration

- Databricks is a collaborative analysis platform where users can share workspaces, clusters and jobs through a single interface.

- It is possible to create shared models in the same real time notebook, reuse data assets, libraries on the same cluster, or reuse/monitor scheduled jobs.

- Databricks supports integration with Github, Bitbucket Cloud & Azure DevOps Services.



https://community.cloud.databricks.com/login.html

Topics

1. What is Machine Learning?
2. Databricks (intro)
3. **Design a pipelines**
    1. How to scrape online job vacanciesBuild our pipeline with Spark
    2. Focus on occupation’s categorization

What are we going to do?

- Collect some job postings
- Create a cluster
- Train a ML model to classify occupations

Topics

1. What is Machine Learning?
2. Databricks (intro)
3. Design a pipelines
    1. **How to scrape online job vacanciesBuild our pipeline with Spark**
    2. Focus on occupation’s categorization

https://www.amazon.jobs/it/search?base_query=&loc_query=

amazonjobs | Search for jobs by title or keyword | Location | Your job application

Filter by

JOB TYPE
- Full Time (47479)
- Part Time (195)
- Seasonal (146)

JOB CATEGORY
- Software Development (12501)
- Solutions Architect (5609)
- Project/Program/Product Management--Non-Tech (3101)
- Operations, IT, & Support Engineering (2953)
- Project/Program/Product Management--Technical (2812)

more

LOCATIONS
- Seattle, Washington, USA (10925)
- Bengaluru, Karnataka, IND (1656)
- Arlington, Virginia, USA (1620)
- New York, New York, USA (1615)

Showing 1 - 10 of 47833 jobs

Sort by: Most relevant

Atendimento ao Cliente -Temporário- Brasil
Posted March 26, 2021
(Updated 40 minutes ago)
BRA | Job ID: SF210055816
O Centro Virtual de Atendimento ao Cliente da Amazon no Brasil está em busca de candidatos com perfis inovadores, dinâmicos e detalhistas com desejo de ajudar a superar as expectativas dos nossos clientes. Os atendentes da Amazon são uma parte fundamental...Read more

Virtual Customer Service Associate – Kolkata, India
Posted March 26, 2021
(Updated about 3 hours ago)
IND, WB, VCC - West Bengal | Job ID: SF210055809
Customer Service Associate-VCS-IndiaAn Amazon Customer Service Associate is a critical part of our mission to deliver timely, accurate and professional customer service to all Amazon customers. This vital position requires an action-oriented, flexible pr...Read more

Delivery Station Liaison - Full Time (40 Hours) - DSX7 - San Antonio, TX, USA
Posted March 25, 2021
(Updated about 1 hour ago)
USA, TX, San Antonio | Job ID: SF210055779
The address for this role, is: 8210 Aviation Landing, San Antonio, TX 78235The schedule for this role, subject to change based on business need, will be: Monday-Friday 10:00AM-7:00PMThis is a Full-Time (40 hours per week) position. The average amount of s...Read more

Delivery Station Liaison - Full Time (40 Hours) - DDX2 - McKinney, TX, USA
Posted March 25, 2021
(Updated about 1 hour ago)
USA, TX, McKinney | Job ID: SF210055778
The address for this role, is: 1398 Industrial Boulevard, McKinney, TX 75069The schedule for this role, subject to change based on business need, will be: Monday-Friday 10:00AM-7:00PMThis is a Full-Time (40 hours per week) position. The average amount of ...Read more

amazonjobs | Search for jobs by title or keyword | Germany | Your job application

Filter by

JOB TYPE

- Full Time (1561)
- Part Time (35)
- Seasonal (1)

JOB CATEGORY

- Fulfillment & Operations Management (277)
- Software Development (149)
- Operations, IT, & Support Engineering (132)
- Solutions Architect (119)
- Sales, Advertising, & Account Management (118)

more

Distance: Mi | Km

5 | 15 | 25 | 35 | 50 | Any

LOCATIONS

- Munich, Bavaria, DEU (460)
- Berlin, Berlin, DEU (314)

Showing 1 - 10 of 1596 jobs

Sort by: Most relevant

Kundenservice im Homeoffice (m/w/d) – Teilzeit (20 Std./Woche)

Posted March 17, 2021 (Updated 9 days ago)

DEU, Standortuebergreifend | Job ID: SF210055424

Rolle: Kundenservice im Homeoffice (m/w/d)Job Typ: 20 Stunden Teilzeit mit Vollzeitstunden in der Hochsaison (Details unten)Ort: Deutschland - bei Dir zu Hause!Amazon VCC GmbHKarl-Liebknecht-Str. 510178 BerlinDeutschlandDeine Herausforderung. Dein Team. D...Read more

Social Media Customer Service Associate (w/m/d) Teilzeit (20 Stunden)

Posted March 16, 2021 (Updated 9 days ago)

DEU, BY, Regensburg | Job ID: SF210055414

Das Social Media Team in Regensburg sucht zum nächstmöglichen Zeitpunkt mehrere Social Media Specialists (m/w/d)Amazon Deutschland Services GmbHIm Gewerbepark D 55 (Main Entrance: D 65)93059 RegensburgDeutschlandDas Social Media Customer Service (SMCS) Pr...Read more

Social Media Customer Service Associate (w/m/d) Vollzeit

Posted March 16, 2021 (Updated 9 days ago)

DEU, BY, Regensburg | Job ID: SF210055413

Das Social Media Team in Regensburg sucht zum nächstmöglichen Zeitpunkt mehrere Social Media Specialists (m/w/d)Amazon Deutschland Services GmbHIm Gewerbepark D 55 (Main Entrance: D 65)93059 RegensburgDeutschlandDas Social Media Customer Service (SMCS) Pr...Read more

Kundenservice im Homeoffice (m/w/d) – Vollzeit (40 Std./Woche)

Posted March 16, 2021 (Updated 9 days ago)

DEU, Standortuebergreifend | Job ID: SF210055410

Dataset

- ~100 Online job ads
  - From amazon.jobs
  - Germany

https://colab.research.google.com/notebooks/intro.ipynb#recent=true

https://pandas.pydata.org/getting_started.html

```
[ ] import pandas as pd
    ds_items = pd.DataFrame(items_details)
    ds_items.set_index("job_id")
    ds_items.head()
```

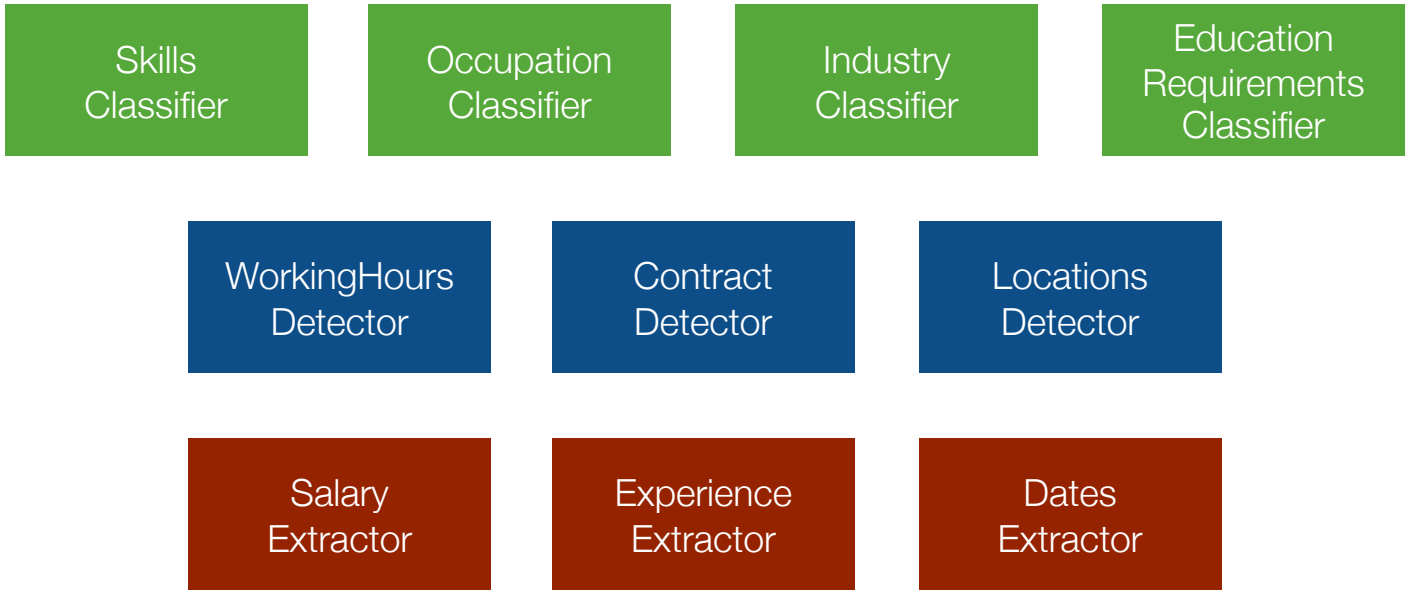
	title	url	location	description	job_id
0	Kundenservice im Homeoffice (m/w/d) – Teilzeit...	https://www.amazon.jobs/en/jobs/SF210055424/ku...	[location]	DESCRIPTION\nRolle: Kundenservice im Homeoffic...	SF210055424
1	Social Media Customer Service Associate (w/m/d...	https://www.amazon.jobs/en/jobs/SF210055414/so...	[location]	DESCRIPTION\nDas Social Media Team in Regensbu...	SF210055414
2	Social Media Customer Service Associate (w/m/d...	https://www.amazon.jobs/en/jobs/SF210055413/so...	[location]	DESCRIPTION\nDas Social Media Team in Regensbu...	SF210055413
3	Kundenservice im Homeoffice (m/w/d) – Vollzeit...	https://www.amazon.jobs/en/jobs/SF210055410/ku...	[location]	DESCRIPTION\nKundenservicemitarbeiter*innen im...	SF210055410
4	Kundenservice (m/w/d) – Berlin – deutsprachige...	https://www.amazon.jobs/en/jobs/SF210054326/ku...	[location]	DESCRIPTION\nKundenservicemitarbeiter*innen / ...	SF210054326

```
ds_items.to_csv('ds_items.csv')
```

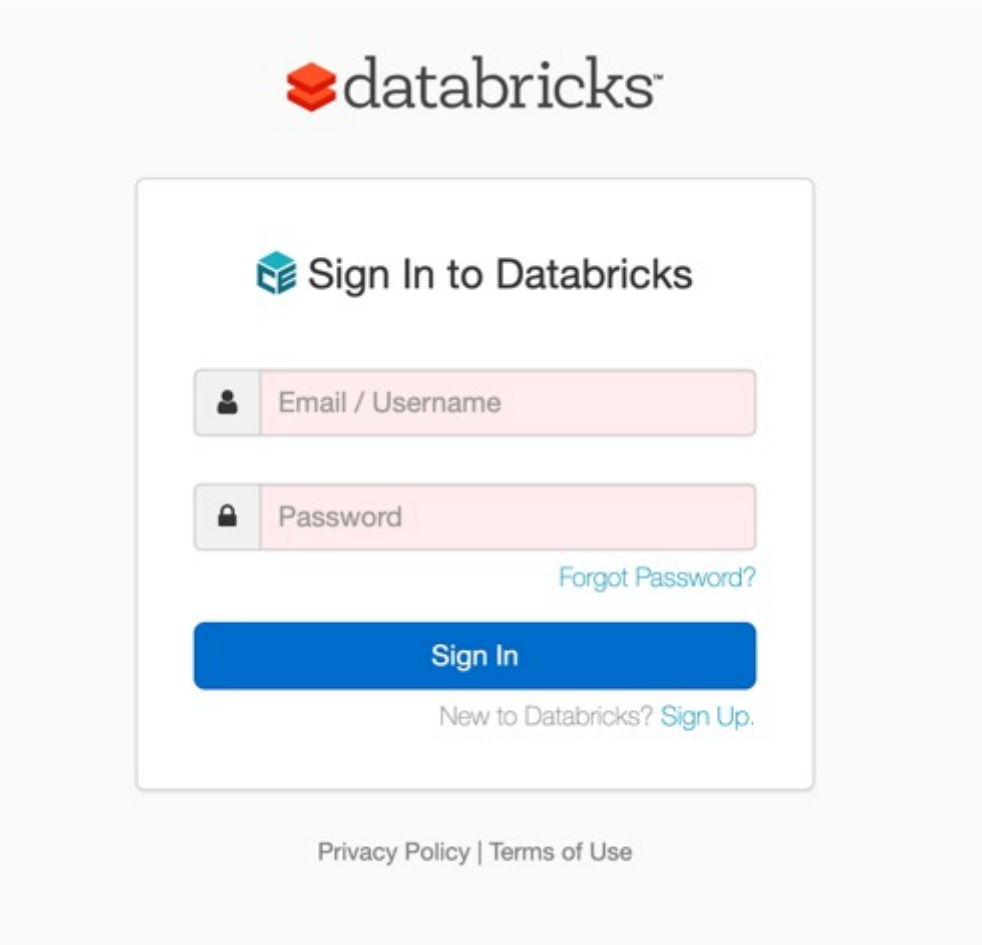
Topics

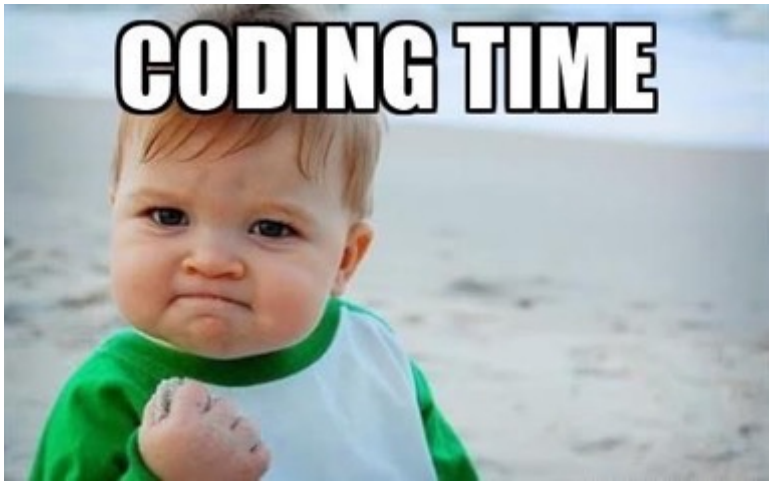
1. What is Machine Learning?
2. Databricks (intro)
3. Design a pipelines
    1. How to scrape online job vacanciesBuild our pipeline with Spark
    2. **Focus on occupation's categorization**

Classification Microservices

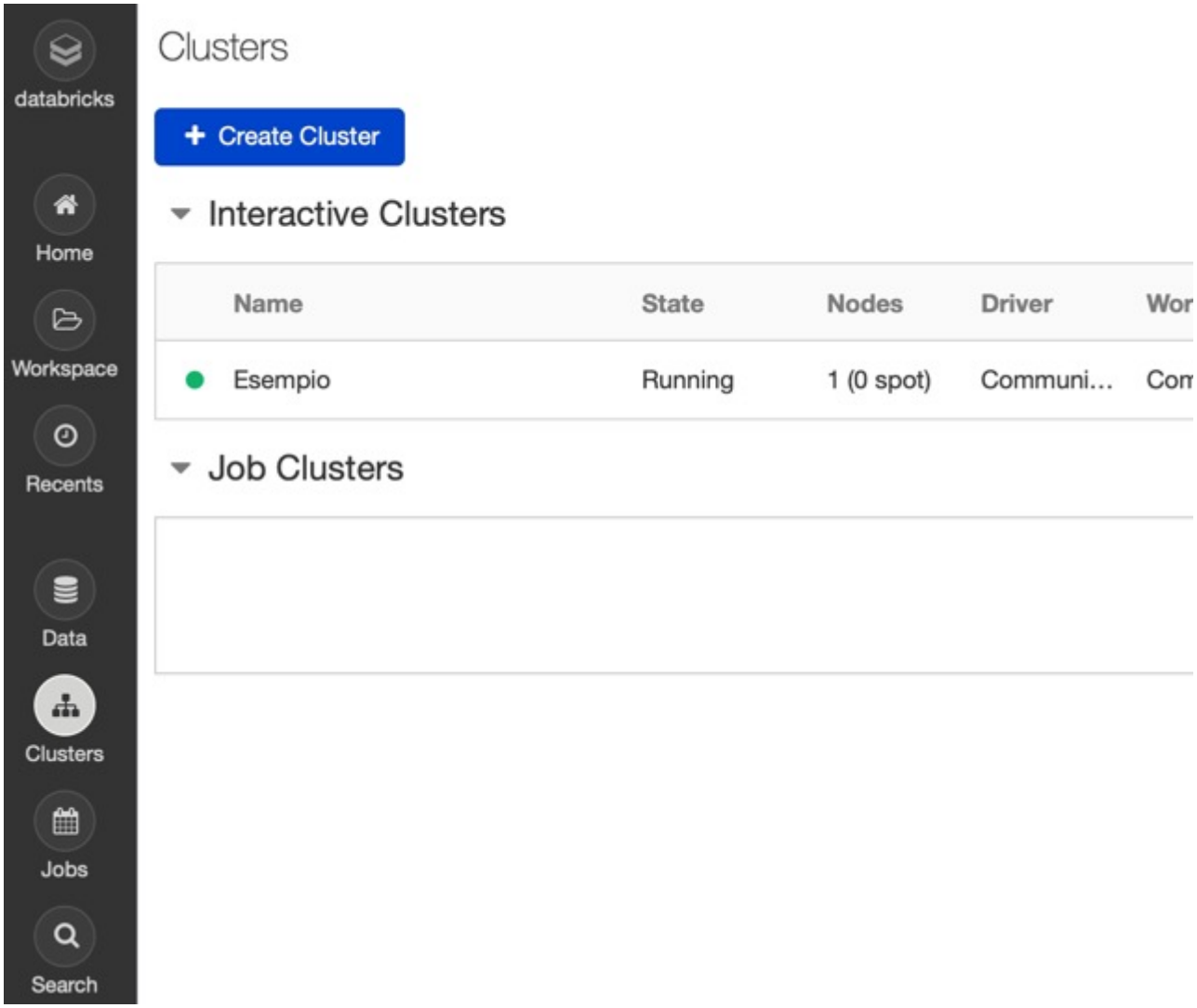


https://community.cloud.databricks.com/login.html

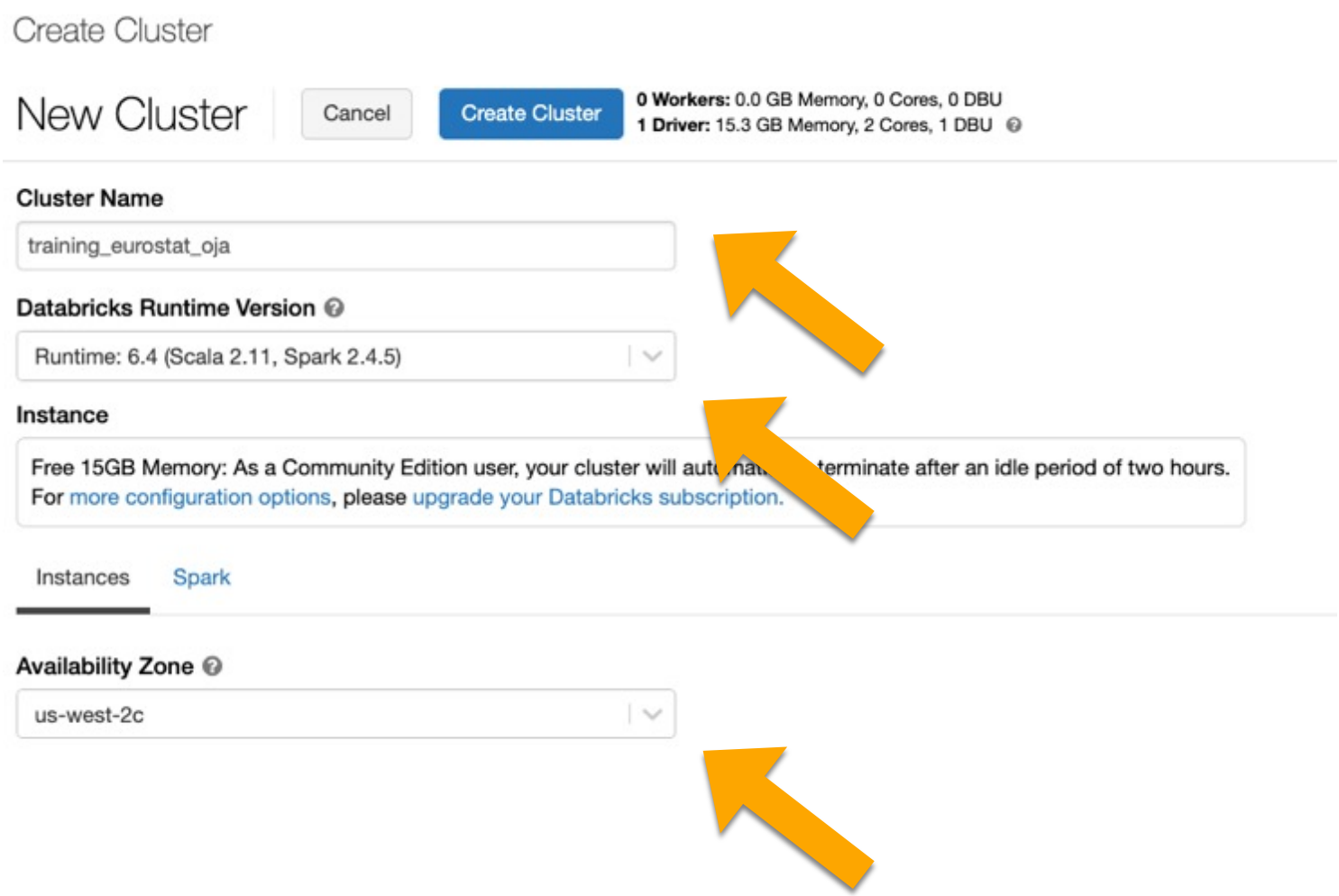




Create a new cluster



Create a new cluster



Create a new cluster

Create Cluster

New Cluster | Cancel | Create Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU

Cluster Name

training_eurostat_oja

Databricks Runtime Version

Runtime: 6.4 (Scala 2.11, Spark 2.4.5)

Instance

Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please upgrade your Databricks subscription.

Instances | Spark

Spark Config

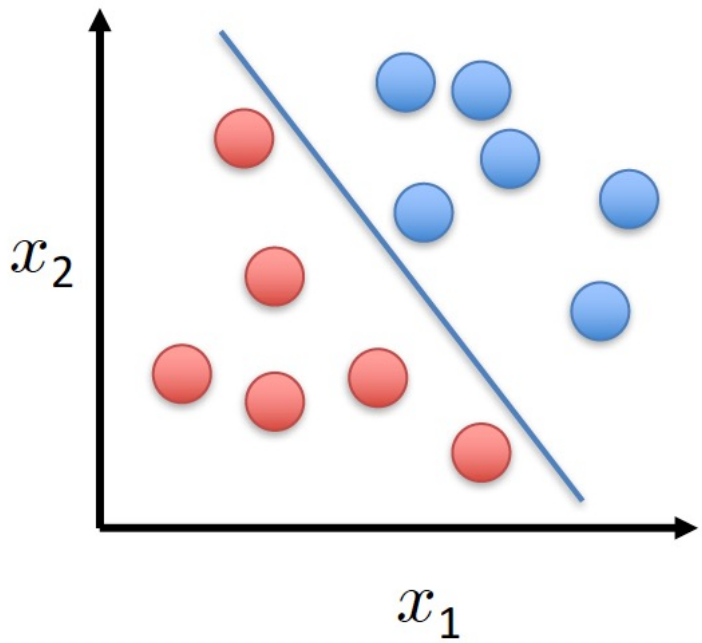
```
spark.mongodb.output.uri
mongodb+srv://admin:1234admin!@lmi.dbru3.mongodb.net/metadata.test
spark.mongodb.input.uri
mongodb+srv://admin:1234admin!@lmi.dbru3.mongodb.net/metadata.test
spark.databricks.delta.preview.enabled true
```

Goals

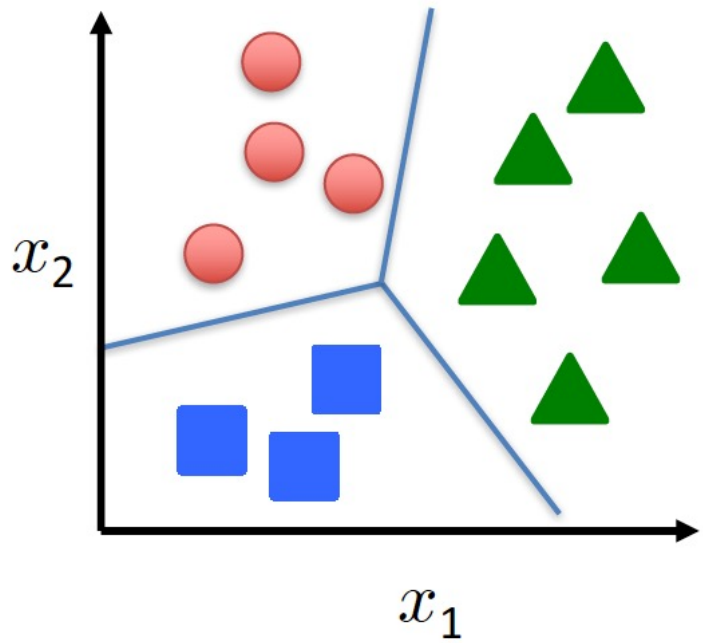
- <u>Classify occupation from title of job ads</u>
- Develop a generic approach valid for all 25 official languages of European Union (but also the co-official languages)
- Reduce the use of gold datasets with the scope to minimize the impact of human errors and ambiguities
    - ~100,000 manually labeled observations for each language
- Design a system that is easily controllable and can be improved in case of misclassification: importance of explainable of outputs

Multi-Class Classification

Binary classification:

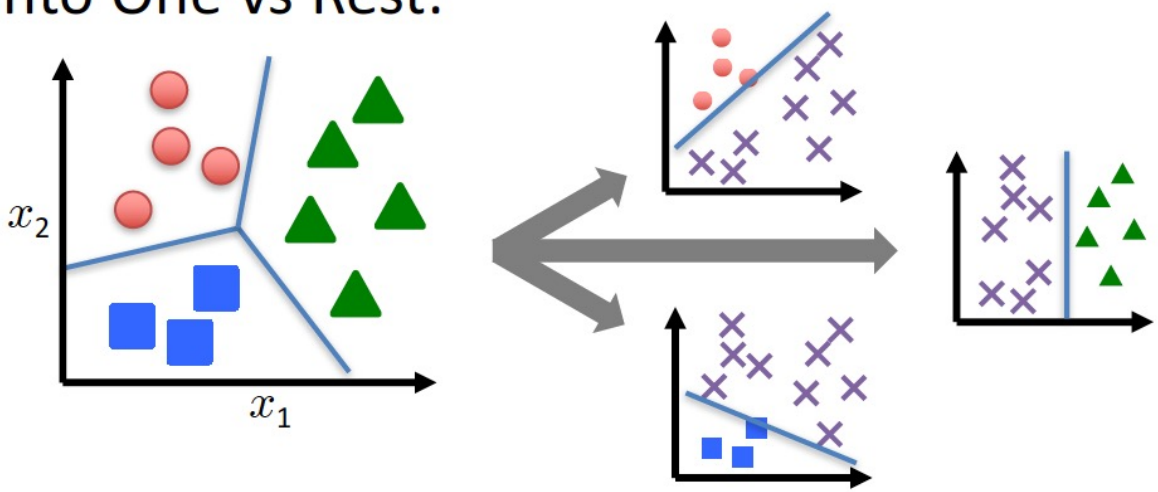


Multi-class classification:



Multi-Class Logistic Regression

Split into One vs Rest:



- Train a logistic regression classifier for each class i to predict the probability that $y = i$ with

$$h_c(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}_c^\top \boldsymbol{x})}{\sum_{c=1}^{C} \exp(\boldsymbol{\theta}_c^\top \boldsymbol{x})}$$

Output of the model
Multi-class model

Job Class: 2512 Software Developer
Job Hier: (251 – IT Profession, 25 -, ….)
Max Prob: 0.98

Predict Probabilties
(for each class)

[
1330: 0.12,
2511: 0.11,
2512: 0.98,
2513: 0.97.
….
]

Prediction path

{
Prediction from text: false
Prediction from metadata: true
Type of model: Ontology / ML
Language of the model: EN
Distance (if appliaed): Equals / Jaccar / Jarowinkler
Distance (metrices): 0..100
Ontology path: software developer
}

The dataset

- 20k online job vacancies
- 4 classes (5k documents for each class)
  - Advertising and marketing professionals
  - Software developers
  - Mathematicians, actuaries and statisticians
  - Industrial and production engineers

Table: esco_en_dataset_csv

esco_en_dataset_csv

Refresh

test (clone)

Schema:

	col_name	data_type	comment
1	title	string	null
2	idesco_level_4	int	null
3	esco_level_4	string	null

Showing all 3 rows.

Sample Data:

	title	idesco_level_4	esco_level_4
1	B93-C04 Softwareentwickler C++ und C#/.NET (m/w)	2512	Software developers
2	Gezocht: Oracle Developer #Freelance #PandS #Jobs #Vacatures (Req:9096–Loc:Bxl)	2512	Software developers
3	Senior (GXP Process Excellence) Engineer	2512	Software developers
4	Software-Entwickler (m/w/d) Buildsystem / Integration	2512	Software developers
5	Business Intelligence Developer	2512	Software developers
6	Microsoft Dynamics NAV Functional Consultant	2512	Software developers

Showing all 20 rows.

ESCO_4Occupation (Python)

test (clone)

Cmd 1

ESCO Occupation Classifier

This notebook (created on Databricks) shows how to train a ML Model with Spark and the use of Spark SQL to clean and prepare the dataset.

The scope is to train a ESCO Occupation Classifier to classify the job vacancies in:

- Advertising and marketing professionals
- Software developers
- Mathematicians, actuaries and statisticians
- Industrial and production engineers

We will use the component of Spark MLIB to transform the input dataset, clean the text, extract the features, train the model and evaluate our results.

Cmd 2

Explore our dataset with SQL

Cmd 3

```
%sql
select count(*) from default.esco_4occupations_csv
```

(2) Spark Jobs

Recap & Keywords

- Data Science project life cicle
- Spark and SparkMlib
- The text mining process on Spark
- How evaluate the model?

Questions?