BEIRUT AI

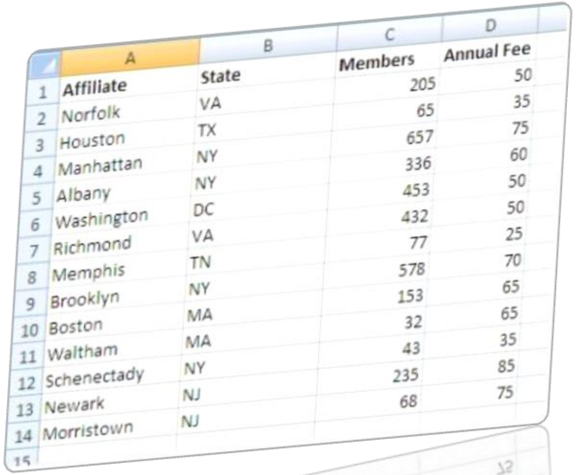
Fully Scalable ML in SQL with Python

Presented by Samer Salameh

Microsoft® SQL Server®

Workshop Outline

- Introduction Data Science
- Predicting Employee Churn
- Preparing SQL Server Environment
- What is SQL?
- SQL Server Management Studio
- Preparing to write SQL
- Machine Learning Services in SQL
- Building AI App

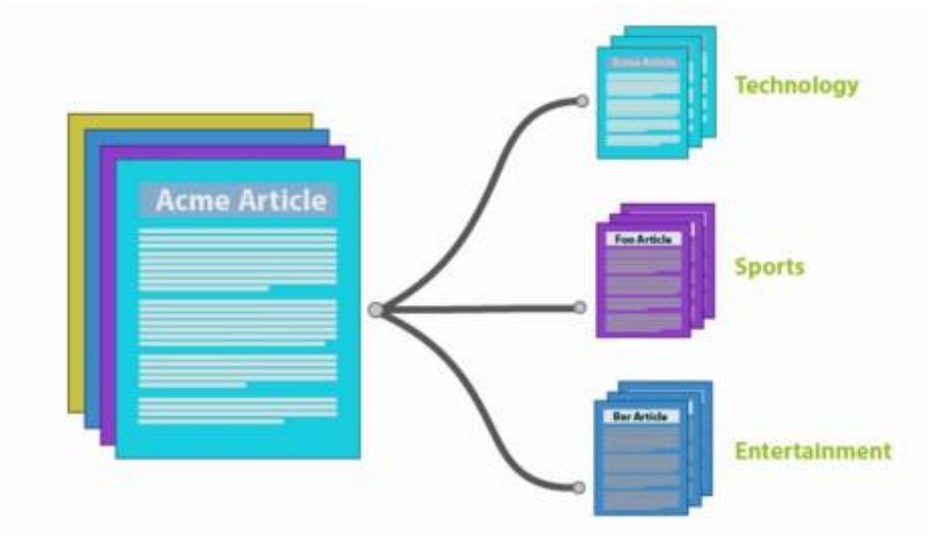


	A	B	C	D
1	Affiliate	State	Members	Annual Fee
2	Norfolk	VA	205	50
3	Houston	TX	65	35
4	Manhattan	NY	657	75
5	Albany	NY	336	60
6	Washington	DC	453	50
7	Richmond	VA	432	50
8	Memphis	TN	77	25
9	Brooklyn	NY	578	70
10	Boston	MA	153	65
11	Waltham	MA	32	65
12	Schenectady	NY	43	35
13	Newark	NJ	235	85
14	Morristown	NJ	68	75
15				

Types of Machine Learning

- Supervised learning
  - Output labels are known
  - Learn the ML model that produces the prediction closest to the output

- Unsupervised Learning
  - Output labels are not known
  - Divides data into clusters of similar data points



VS

Types of Machine Learning

- Reinforcement Learning
  - An agent interacts with an environment and performs action
  - Learns through experience (reward mechanism)

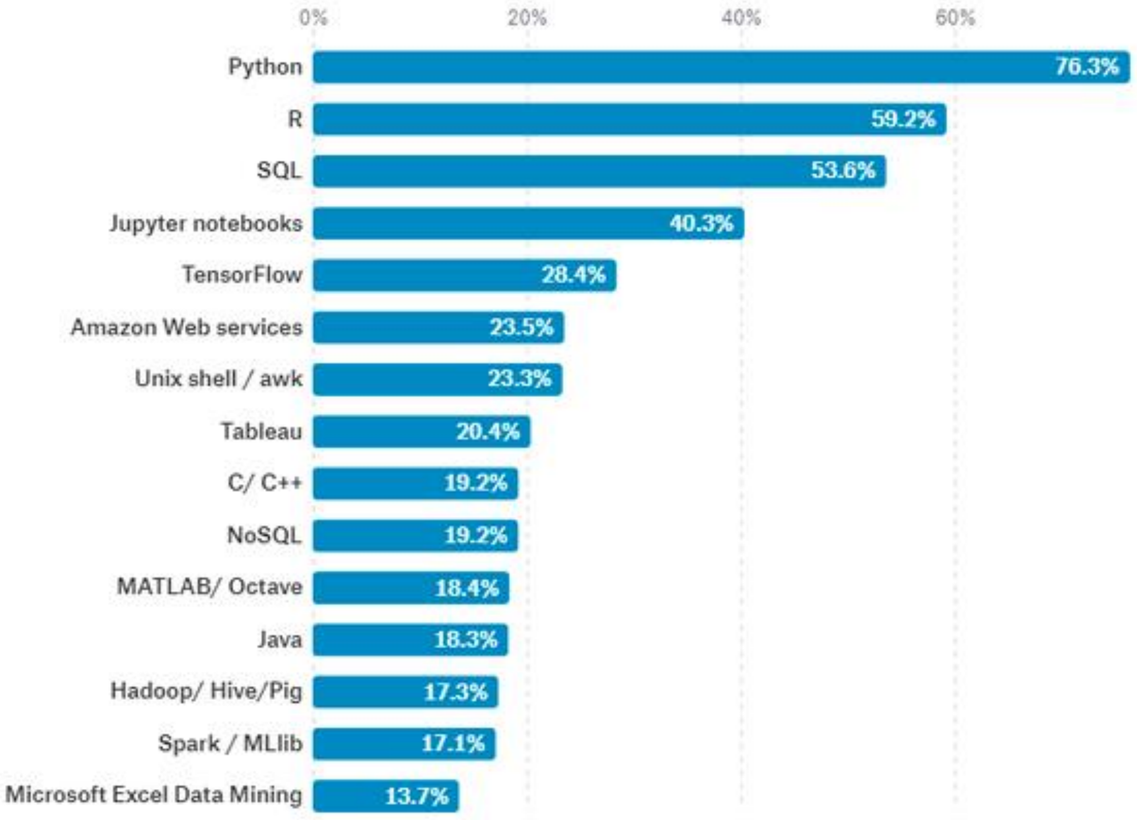
DEEP REINFORCEMENT LEARNING
IN PACMAN

TYCHO VAN DER OUDERAA

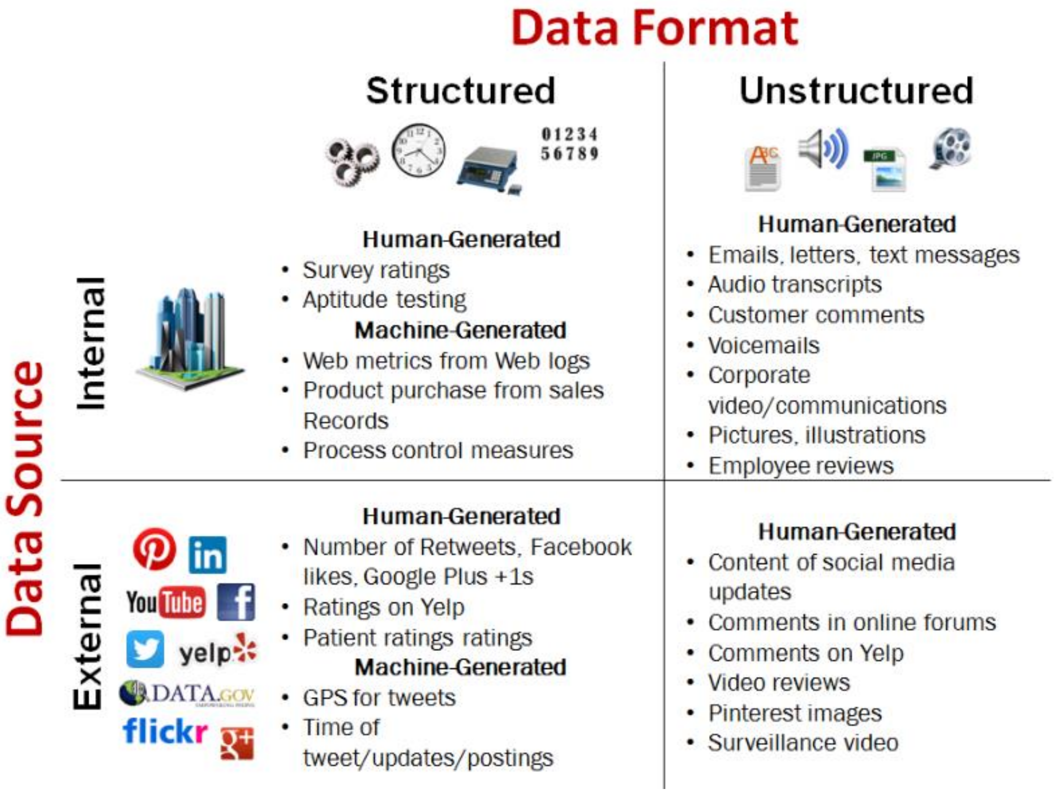
Jupyter Notebook

- Open-source web application
- To create and share documents that contain live code,equations,visualizations and narrative text.
- Uses include
  - data cleaning and transformation
  - numerical simulation
  - statistical modeling
  - data visualization
  - machine learning
  - and much more

Top Data Science Technologies



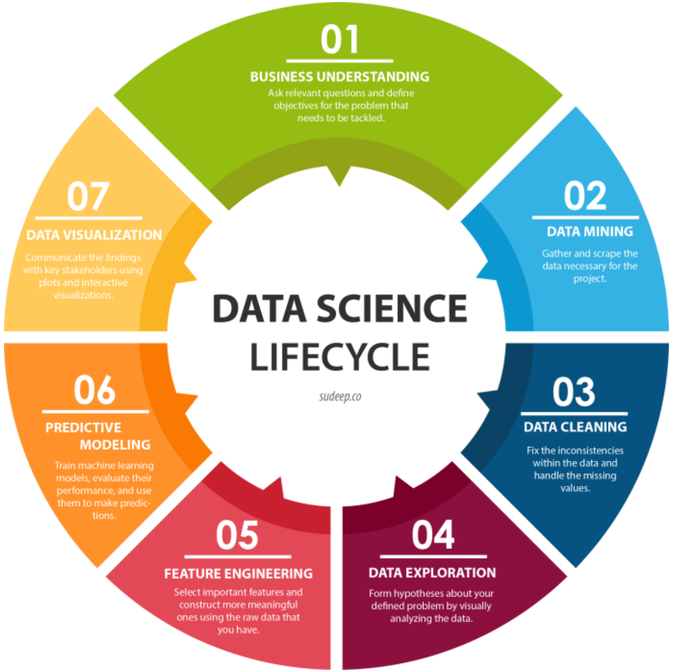
Data Definition Framework



Data Format

Data Source		Structured	Unstructured
Internal		**Human-Generated** • Survey ratings • Aptitude testing **Machine-Generated** • Web metrics from Web logs • Product purchase from sales Records • Process control measures	**Human-Generated** • Emails, letters, text messages • Audio transcripts • Customer comments • Voicemails • Corporate video/communications • Pictures, illustrations • Employee reviews
External		**Human-Generated** • Number of Retweets, Facebook likes, Google Plus +1s • Ratings on Yelp • Patient ratings ratings **Machine-Generated** • GPS for tweets • Time of tweet/updates/postings	**Human-Generated** • Content of social media updates • Comments in online forums • Comments on Yelp • Video reviews • Pinterest images • Surveillance video

Data Science Life Cycle



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

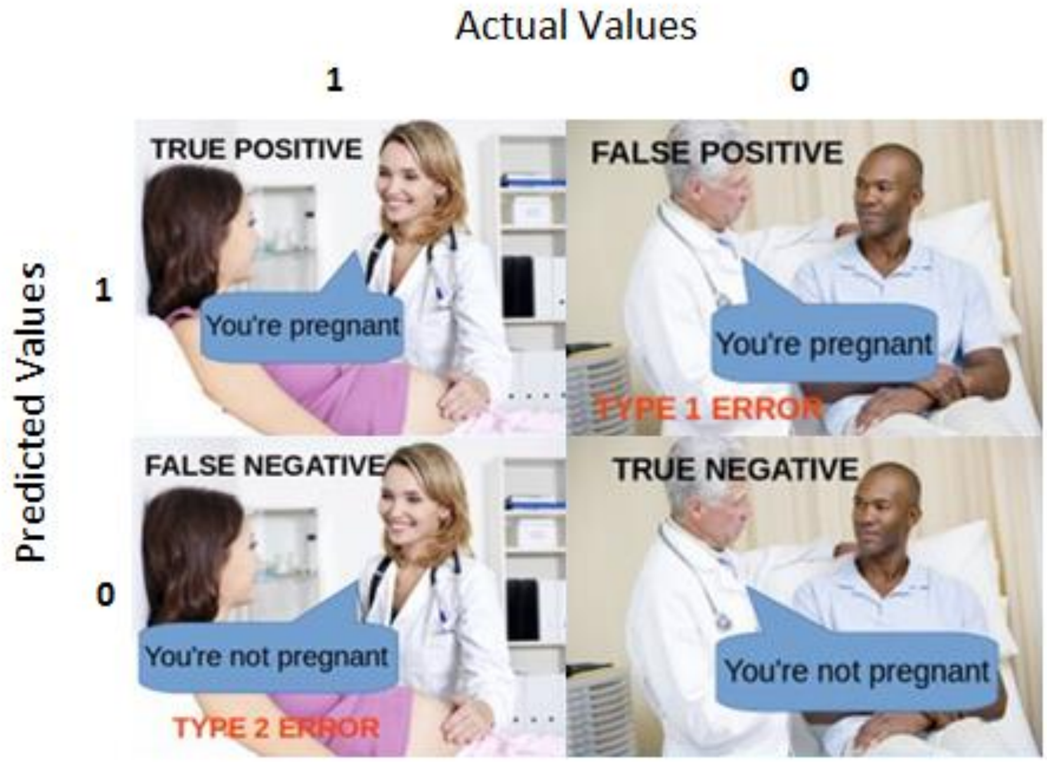
06 PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

Machine Learning Confusion Matrix



PRECISION, RECALL & F1

PRECISION, RECALL AND F1
Uses POSITIVES & NEGATIVE to measure a model’s ACCURACY when making predictions

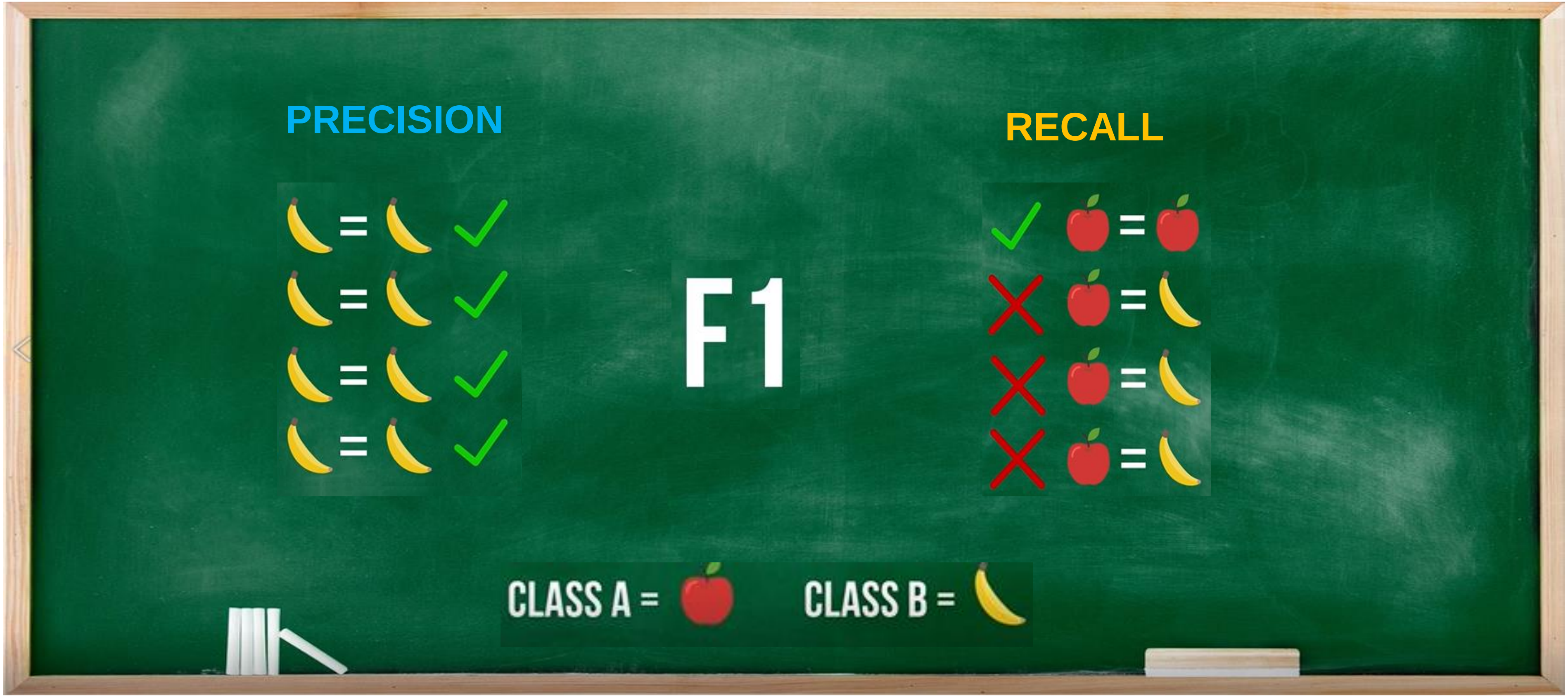
PRECISION
RECALL
F1

ACCURACY

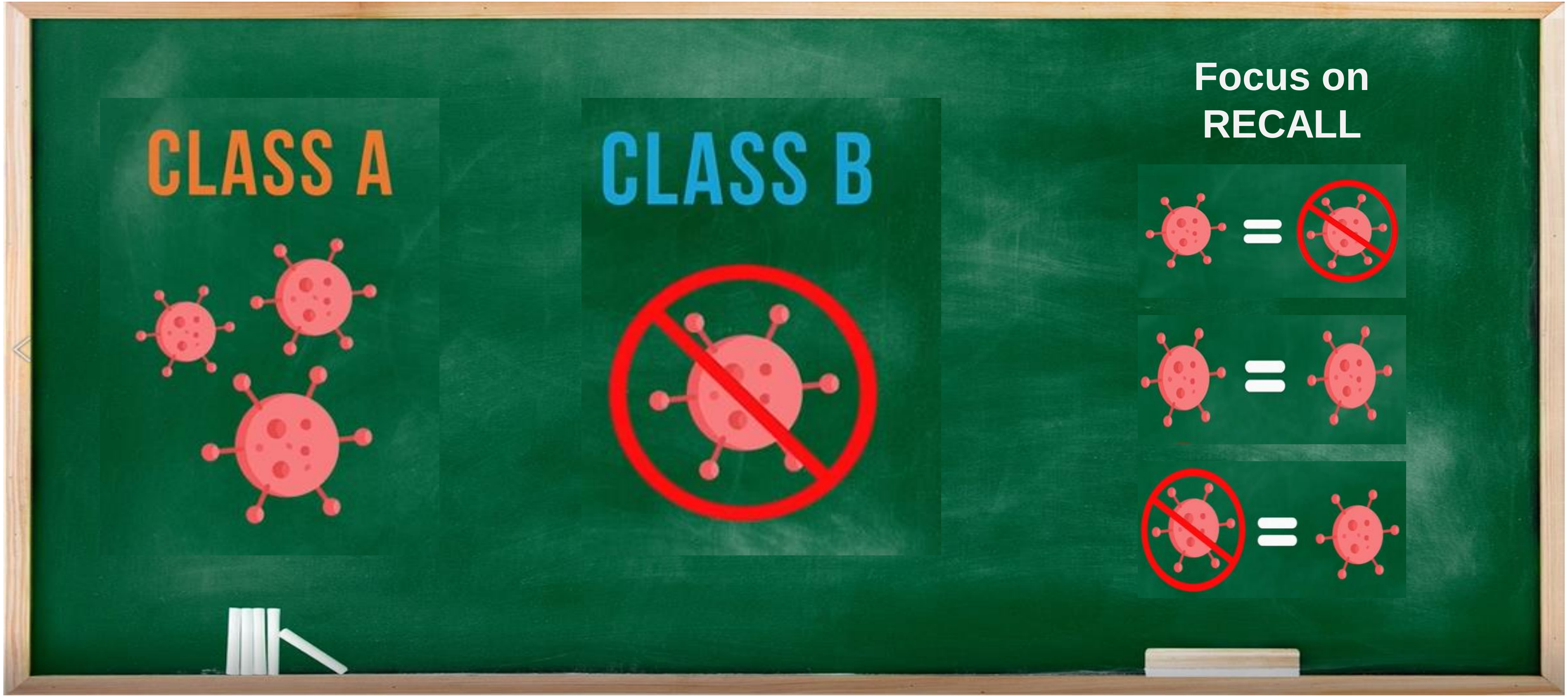
+ +

- -

PRECISION, RECALL & F1



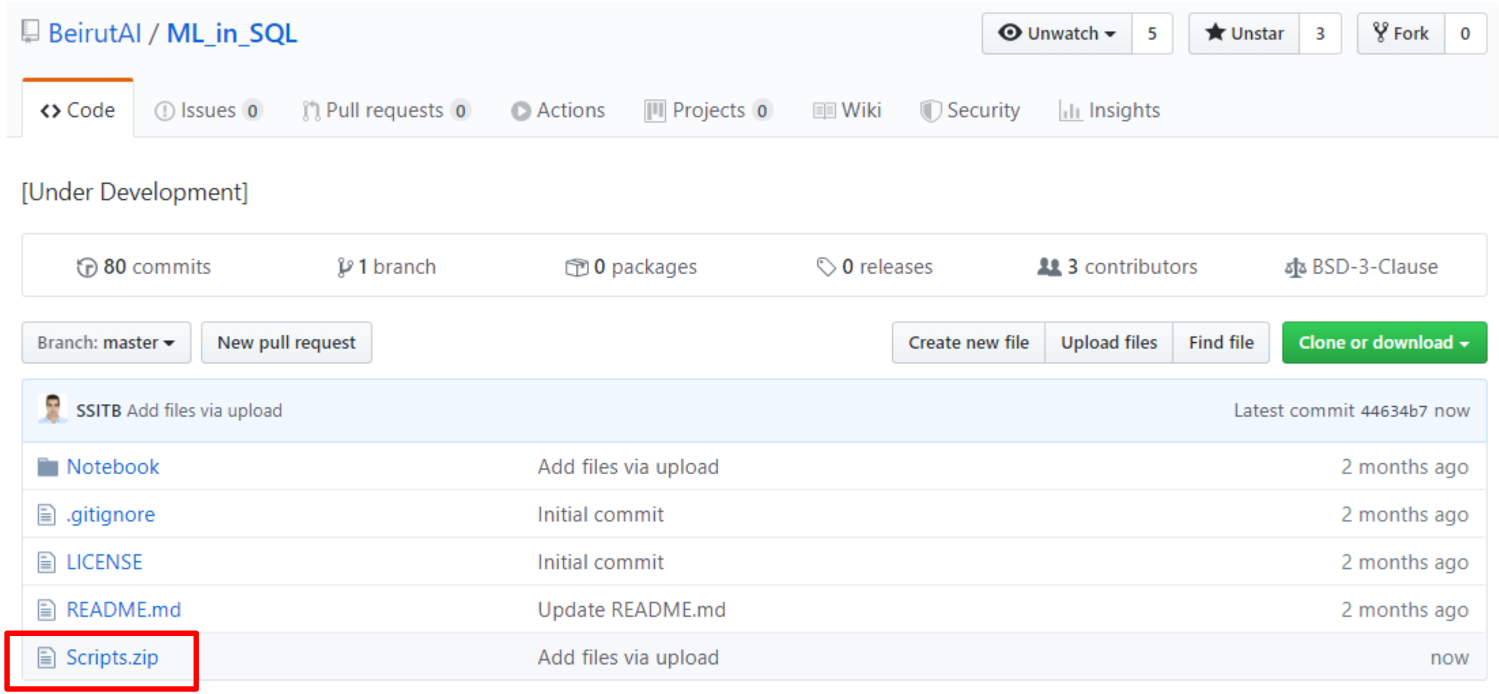
PRECISION, RECALL & F1



Focus on RECALL

Download Scripts

http://github.com/BeirutAI/ML_in_SQL



BeirutAI / ML_in_SQL

[Under Development]

80 commits	1 branch	0 packages	0 releases	3 contributors	BSD-3-Clause

SSITB Add files via upload — Latest commit 44634b7 now

Notebook	Add files via upload	2 months ago
.gitignore	Initial commit	2 months ago
LICENSE	Initial commit	2 months ago
README.md	Update README.md	2 months ago
Scripts.zip	Add files via upload	now

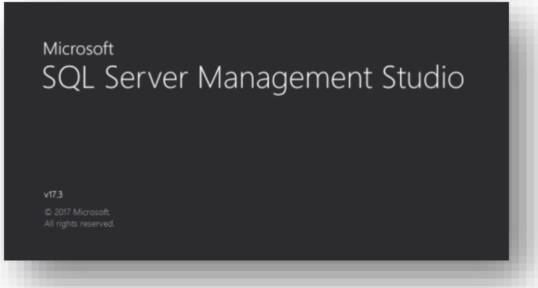
What are databases?

> *A usually large collection of data organized especially for rapid search and retrieval.*

- Holds data
- Organizes data
- Retrieve/Search data through DBMS

SQL Server Management Studio (SSMS)

- SQL Server Management Studio (SSMS) is a powerful graphical DB management tool
  - Administrate databases (create, modify, backup / restore DB)
  - Create and modify Entity Relationship (E/R) diagrams
  - View / modify table data and other DB objects
  - Execute SQL queries
  - Free and easy to use tool
  - Works with all SQL Server versions
  - And much more



Microsoft
SQL Server Management Studio

v17.3

© 2017 Microsoft.
All rights reserved.

Structured and unstructured data

Structured: database schema

- Relational database

Semi-structured

- JSON

```
{ "key": "value"}
```

Unstructured: schemaless, more like files

- Videos, photos

SQL and NoSQL

SQL

- Tables
- Database schema
- Relational databases

MySQL

PostgreSQL

NoSQL

- Non-relational databases
- Structured or unstructured
- Key-value stores (e.g. caching)
- Document DB (e.g. JSON objects)

redis mongoDB

What is SQL

SQL

Structured Query Language

Storage

Access and manipulate

Target

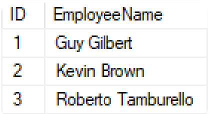
IBM

Developed by

Like All relational database management system

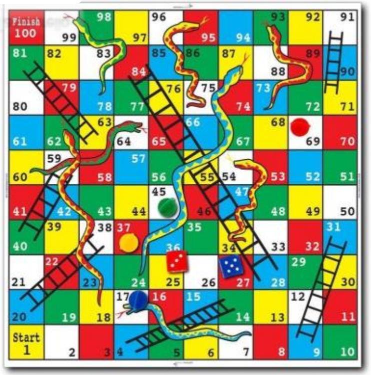
Preparing to write SQL

Easy to learn



ID	EmployeeName
1	Guy Gilbert
2	Kevin Brown
3	Roberto Tamburello

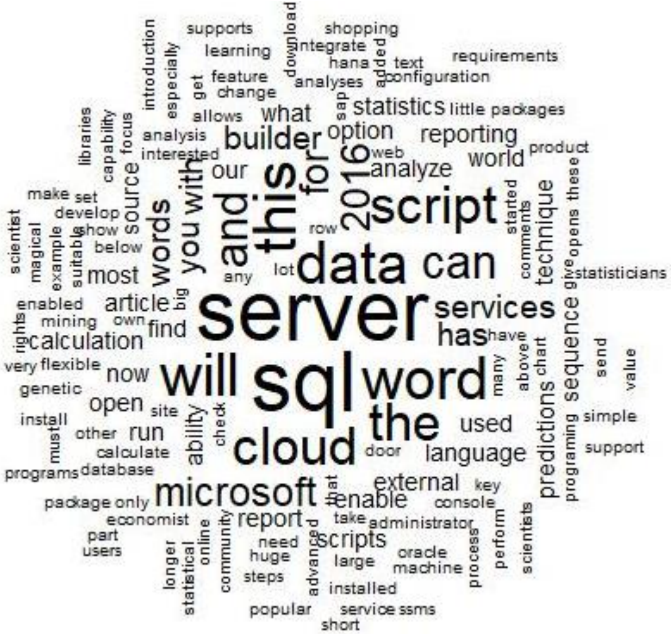
Result



```
Select
    ID,
    EmployeeName
From DimEmployee
```

```
Select
    *
  From DimEmployee
```

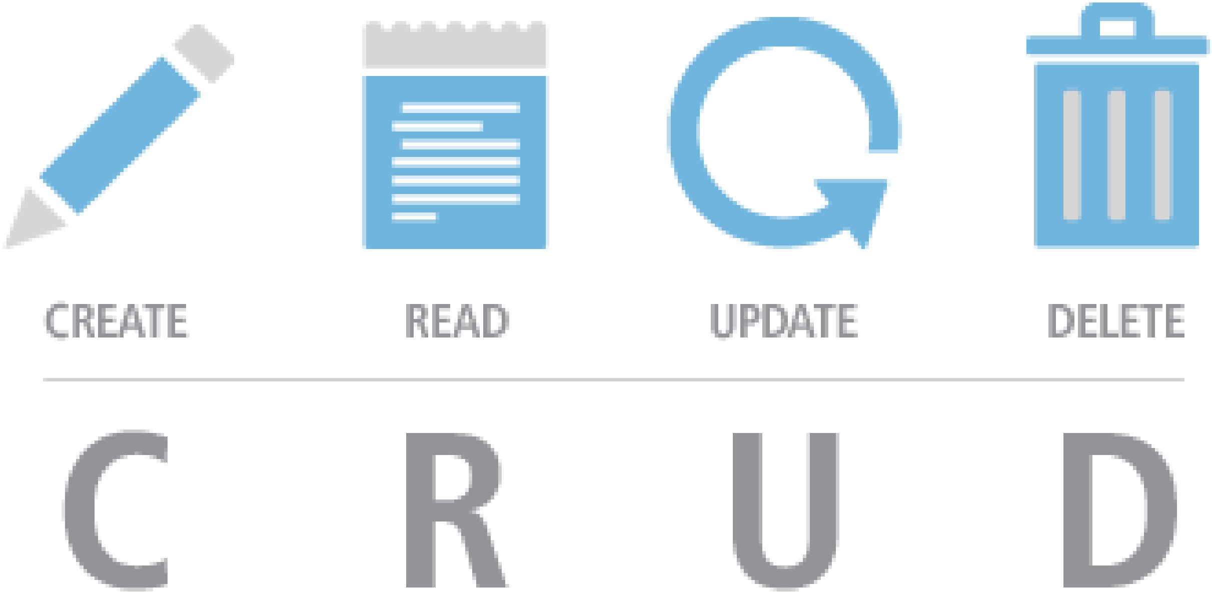
Several SQL dialects exist



Basic command Vocabulary less than 100 words

What SQL can Do?

SQL



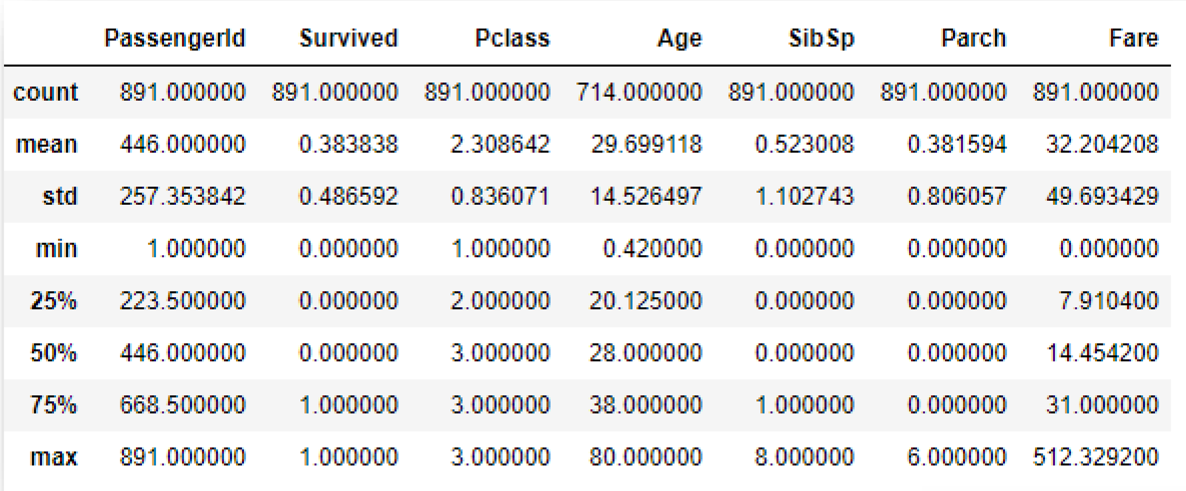
Machine Learning Already in Your Database

- Open database
- Think about use case
- Define business problems
- Extract , Transform, Load (ETL)
- Data wrangling
- Production scripts

Data Engineering

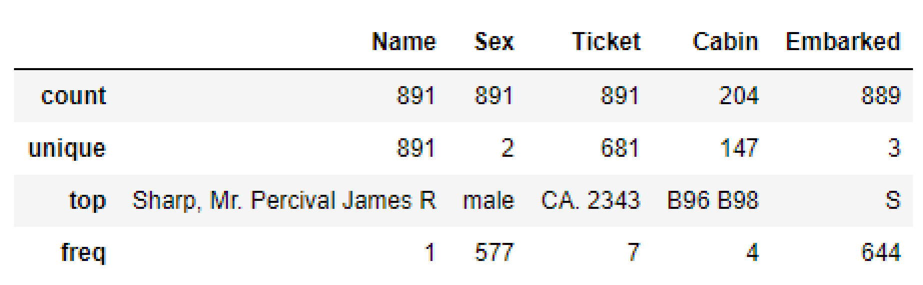
DEMO

Descriptive Data



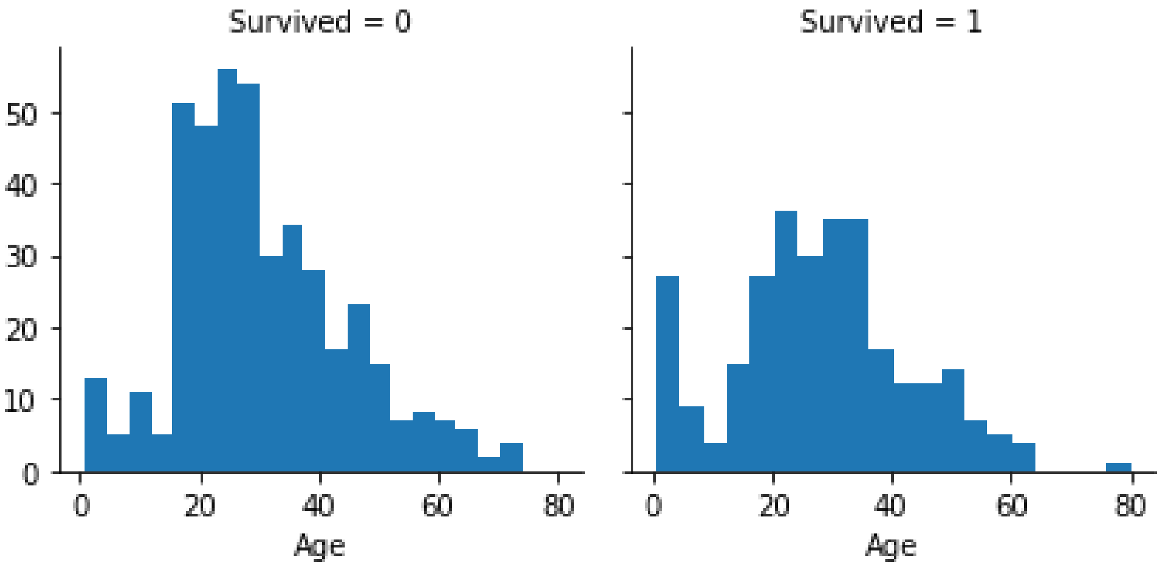
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

- Passengers
- Fares variance (8$ → 512$)
- Males 65%
- Cabin

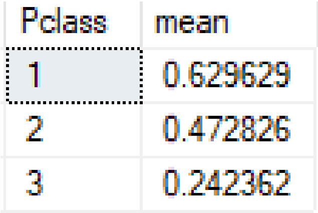


	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Sharp, Mr. Percival James R	male	CA. 2343	B96 B98	S
freq	1	577	7	4	644

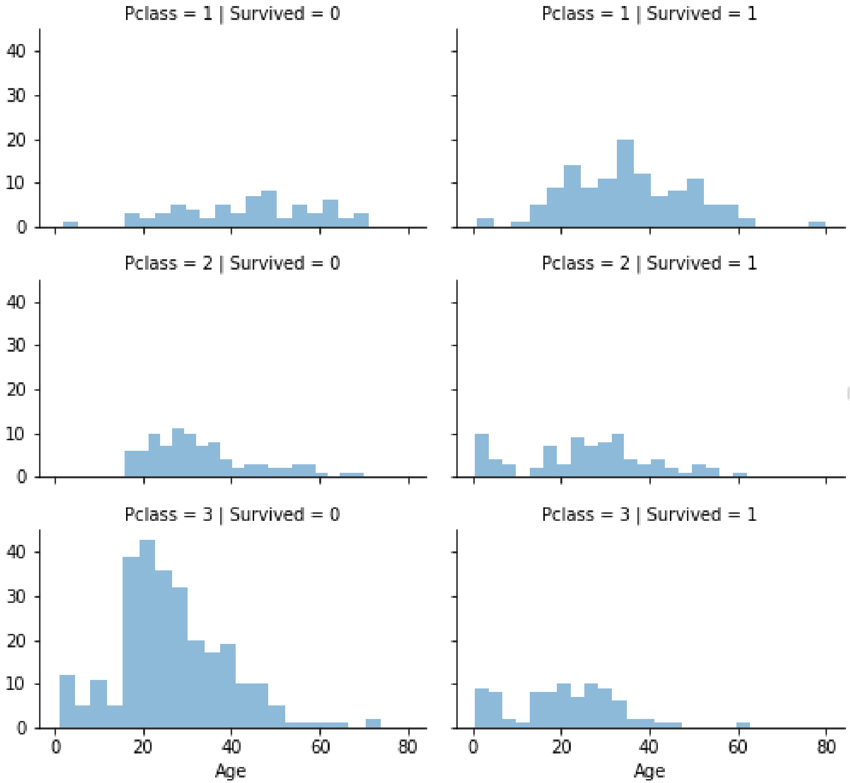
Analyze by visualizing data

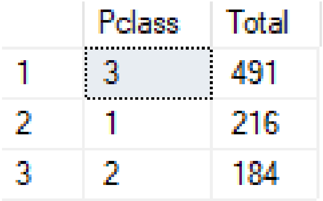


Analyze by visualizing data



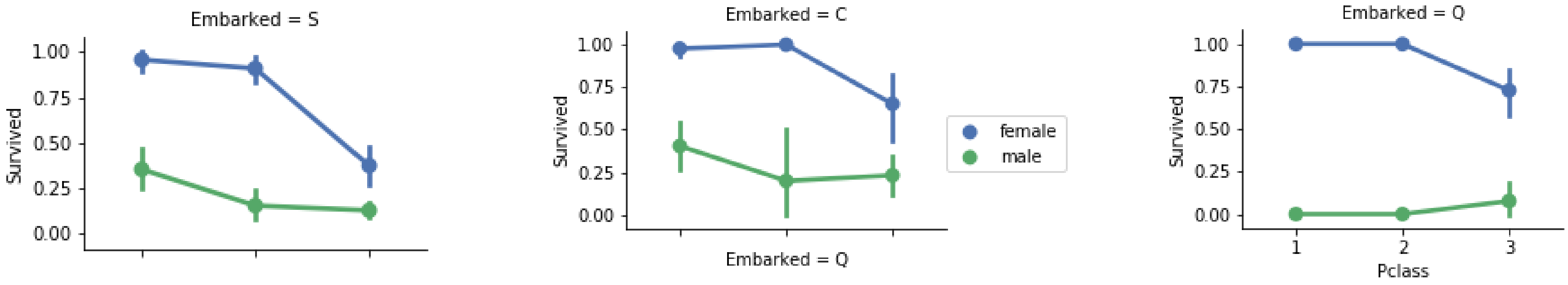
Pclass	mean
1	0.629629
2	0.472826
3	0.242362



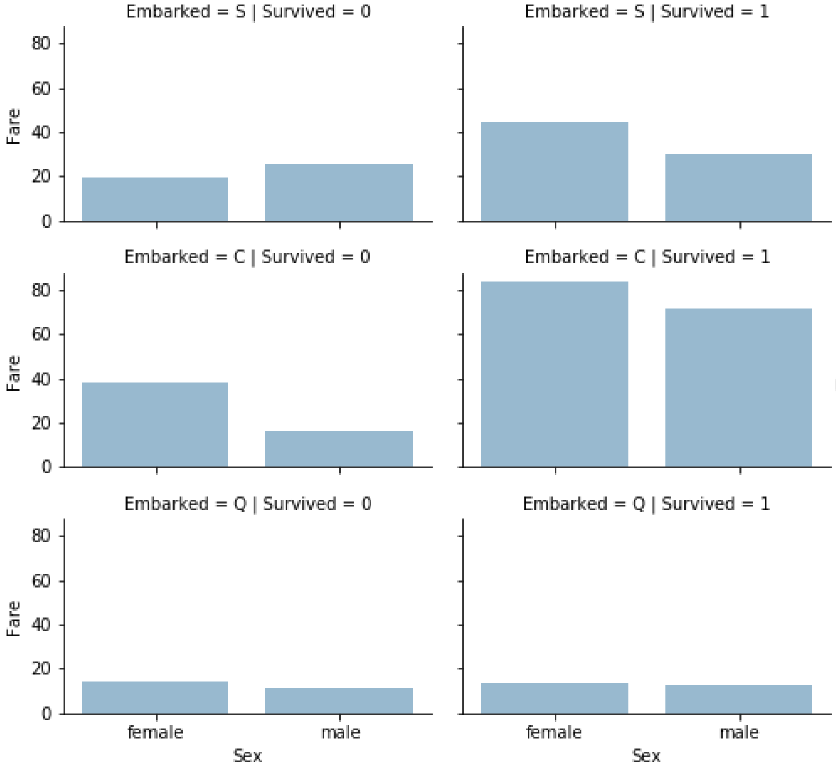


	Pclass	Total
1	3	491
2	1	216
3	2	184

Analyze by visualizing data



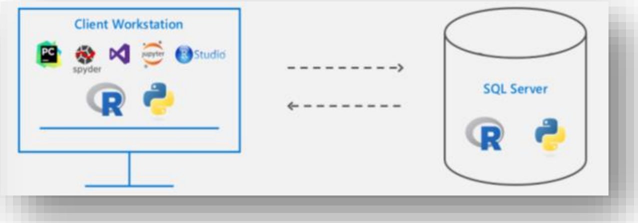
Analyze by visualizing data



Machine Learning Services

Rest API using Flask

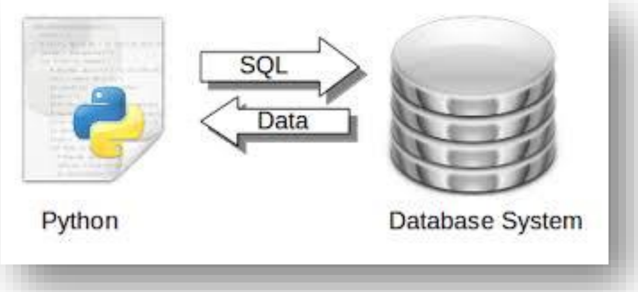
Why Python in SQL



Is it SQL in Python ?

I can use SQLAlchemy ?

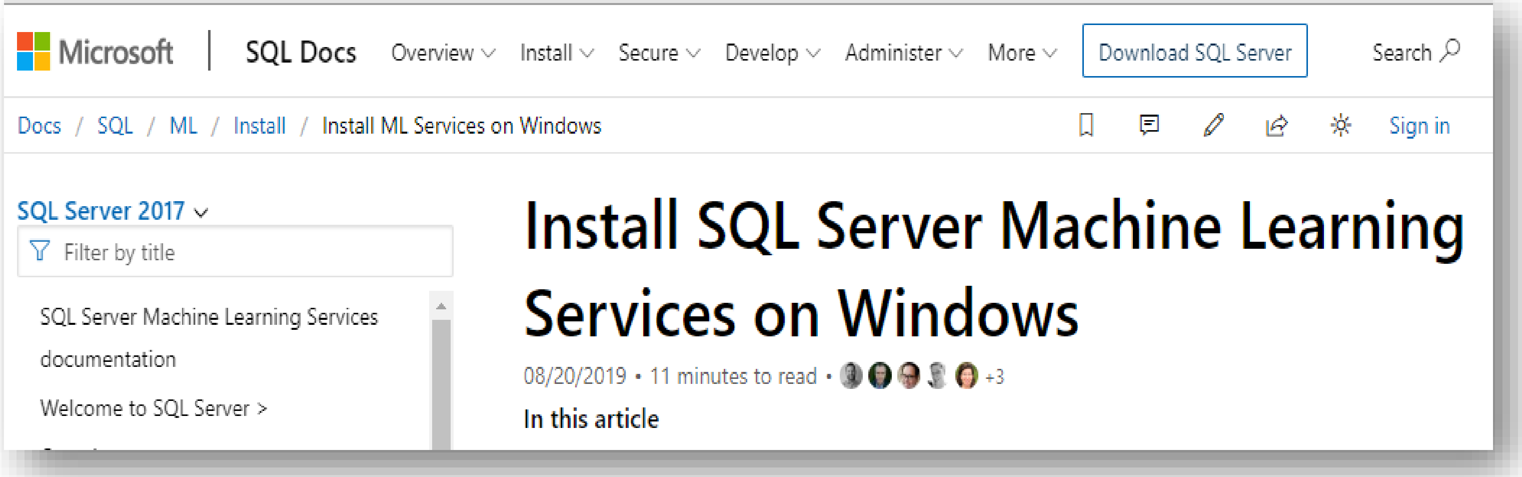
Run external script



Scalability & Speed

Using Python Inside SQL

- Eliminate data movement
- Easily operate Python code inside SQL
- Achieve enterprise grade performance and scale (much faster mechanism than ODBC)



Install SQL Server Machine Learning Services on Windows

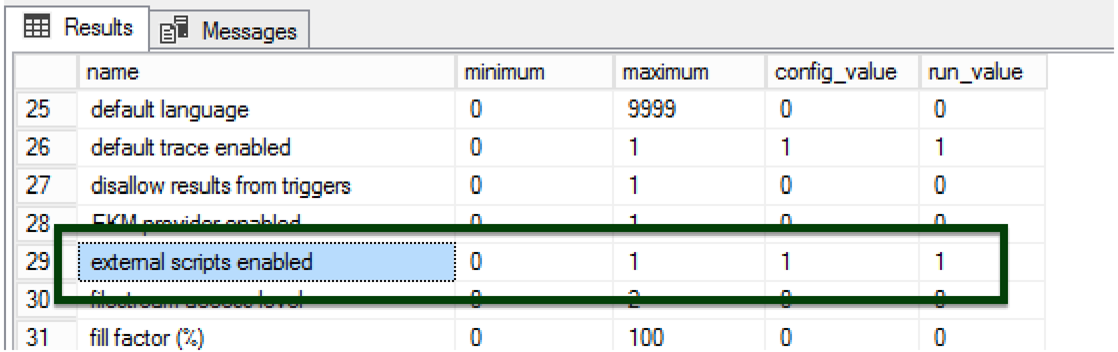
08/20/2019 • 11 minutes to read • +3

In this article

Configuration

- To enable SQL Instance to run Python scripts:

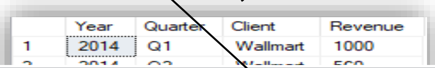
```
sp_configure


EXEC sp_configure 'external scripts enabled', 1
RECONFIGURE WITH OVERRIDE
```



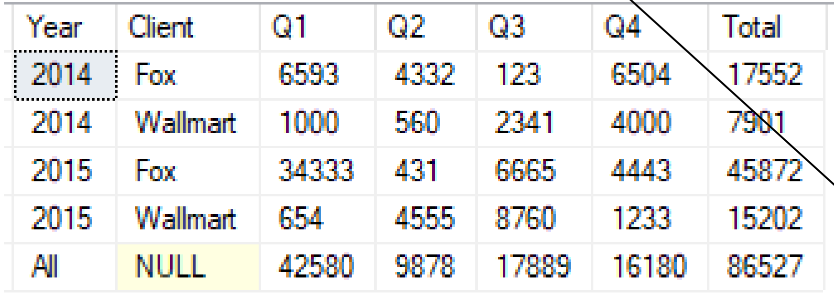
	name	minimum	maximum	config_value	run_value
25	default language	0	9999	0	0
26	default trace enabled	0	1	1	1
27	disallow results from triggers	0	1	0	0
28	EKM provider enabled	0	1	0	0
29	external scripts enabled	0	1	1	1
30	filestream access level	0	2	0	0
31	fill factor (%)	0	100	0	0

Execute External Script

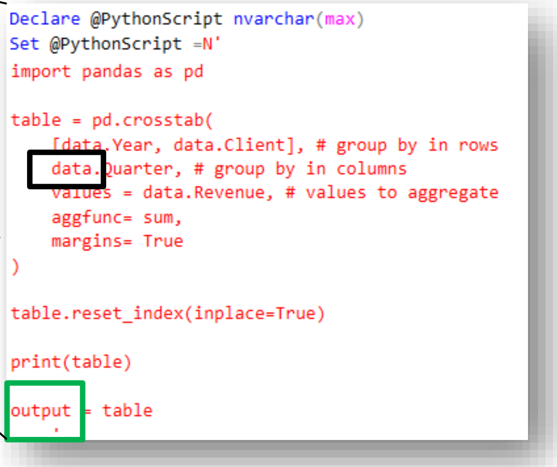
```
EXECUTE sp_execute_external_script
        @language = N'Python'
      , @script = @PythonScript
      , @input_data_1 = N'SELECT CONVERT(VARCHAR, Year) AS Year, Quarter, Client,
Revenue FROM Sales;'
      , @input_data_1_name = N'data'
      , @output_data_1_name = N'output'
WITH RESULT SETS ((
            Year NVARCHAR(10),
            Client NVARCHAR(10),
            Q1 INT,
            Q2 INT,
            Q3 INT,
            Q4 INT,
            Total INT
));
```



	Year	Quarter	Client	Revenue
1	2014	Q1	Wallmart	1000



Year	Client	Q1	Q2	Q3	Q4	Total
2014	Fox	6593	4332	123	6504	17552
2014	Wallmart	1000	560	2341	4000	7901
2015	Fox	34333	431	6665	4443	45872
2015	Wallmart	654	4555	8760	1233	15202
All	NULL	42580	9878	17889	16180	86527



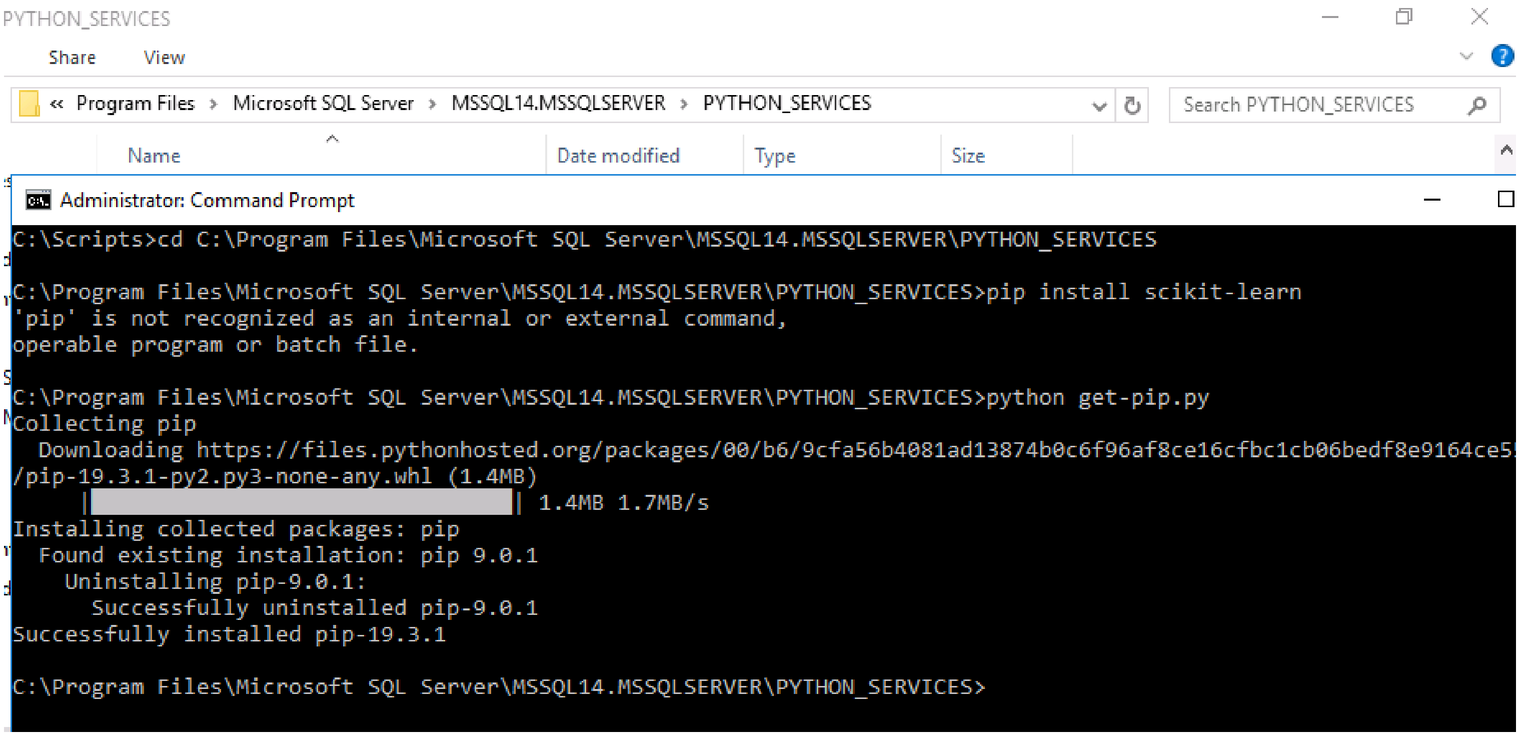
```
Declare @PythonScript nvarchar(max)
Set @PythonScript =N'
import pandas as pd

table = pd.crosstab(
    [data.Year, data.Client], # group by in rows
    data.Quarter, # group by in columns
    values = data.Revenue, # values to aggregate
    aggfunc= sum,
    margins= True
)

table.reset_index(inplace=True)

print(table)

output = table
'
```

Machine Learning

DEMO

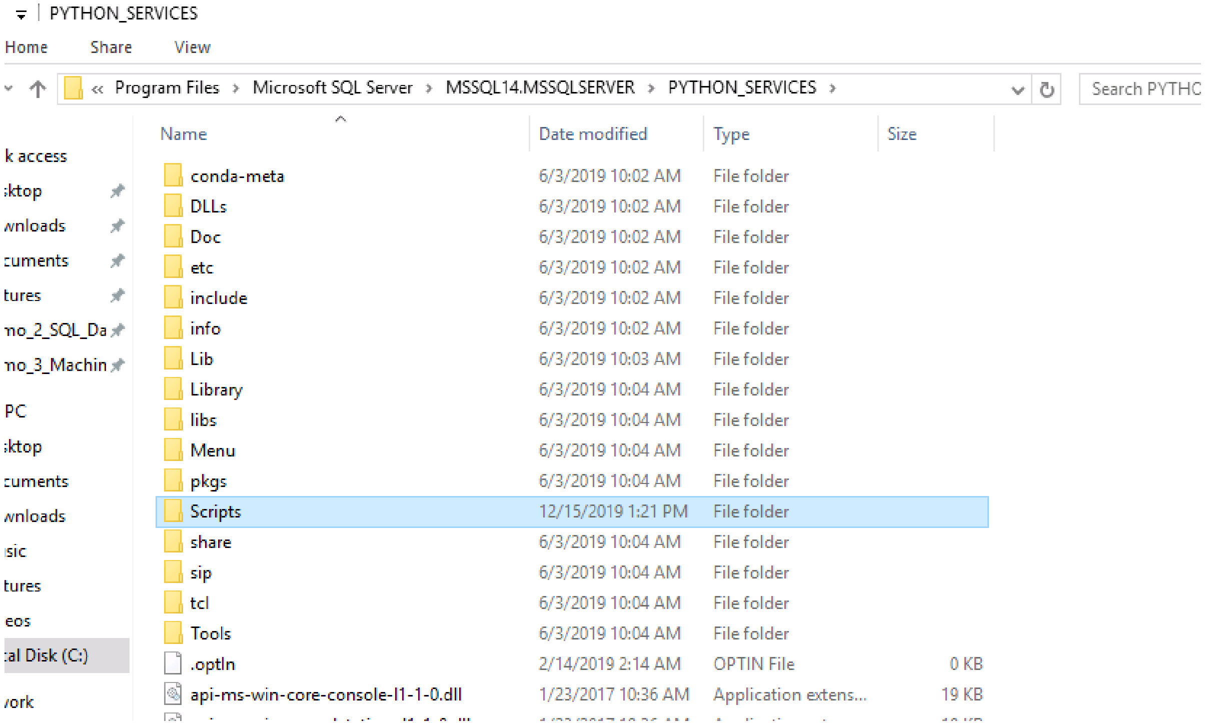
Microsoft® SQL Server®

Install Libraries


PYTHON_SERVICES
Share
View
« Program Files › Microsoft SQL Server › MSSQL14.MSSQLSERVER › PYTHON_SERVICES
Search PYTHON_SERVICES
Name
Date modified
Type
Size
Administrator: Command Prompt

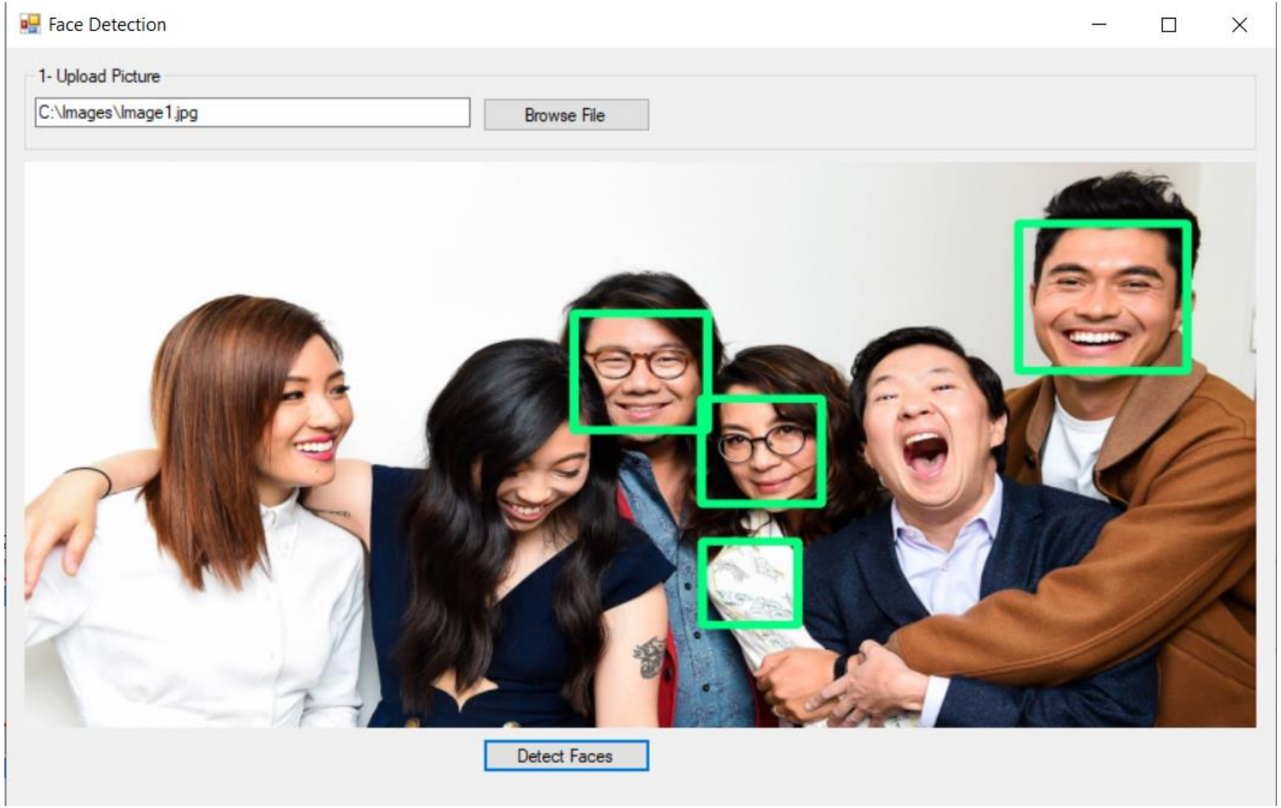
```
C:\Scripts>cd C:\Program Files\Microsoft SQL Server\MSSQL14.MSSQLSERVER\PYTHON_SERVICES

C:\Program Files\Microsoft SQL Server\MSSQL14.MSSQLSERVER\PYTHON_SERVICES>pip install scikit-learn
'pip' is not recognized as an internal or external command,
operable program or batch file.

C:\Program Files\Microsoft SQL Server\MSSQL14.MSSQLSERVER\PYTHON_SERVICES>python get-pip.py
Collecting pip
  Downloading https://files.pythonhosted.org/packages/00/b6/9cfa56b4081ad13874b0c6f96af8ce16cfbc1cb06bedf8e9164ce5
/pip-19.3.1-py2.py3-none-any.whl (1.4MB)
     |████████████████████████████████| 1.4MB 1.7MB/s
Installing collected packages: pip
  Found existing installation: pip 9.0.1
    Uninstalling pip-9.0.1:
      Successfully uninstalled pip-9.0.1
Successfully installed pip-19.3.1

C:\Program Files\Microsoft SQL Server\MSSQL14.MSSQLSERVER\PYTHON_SERVICES>
```

Install Libraries



Name	Date modified	Type	Size
conda-meta	6/3/2019 10:02 AM	File folder	
DLLs	6/3/2019 10:02 AM	File folder	
Doc	6/3/2019 10:02 AM	File folder	
etc	6/3/2019 10:02 AM	File folder	
include	6/3/2019 10:02 AM	File folder	
info	6/3/2019 10:02 AM	File folder	
Lib	6/3/2019 10:03 AM	File folder	
Library	6/3/2019 10:04 AM	File folder	
libs	6/3/2019 10:04 AM	File folder	
Menu	6/3/2019 10:04 AM	File folder	
pkgs	6/3/2019 10:04 AM	File folder	
Scripts	12/15/2019 1:21 PM	File folder	
share	6/3/2019 10:04 AM	File folder	
sip	6/3/2019 10:04 AM	File folder	
tcl	6/3/2019 10:04 AM	File folder	
Tools	6/3/2019 10:04 AM	File folder	
.optIn	2/14/2019 2:14 AM	OPTIN File	0 KB
api-ms-win-core-console-l1-1-0.dll	1/23/2017 10:36 AM	Application extens...	19 KB

Build AI App in 10min



C:\Images\Image1.jpg

Who am I?

Samer Salameh

Senior Data Engineer & Business Analyst
Data Scientist