



**הטכניון**  
מכון טכנולוגי  
לישראל

# בינה מלאכותית ולמידת מכונה

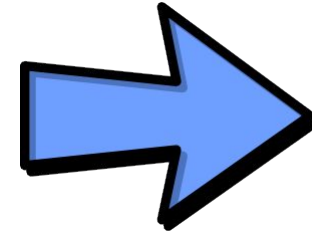
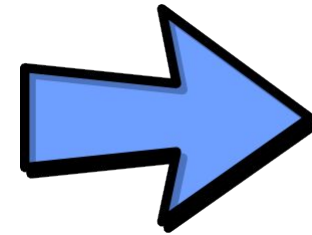
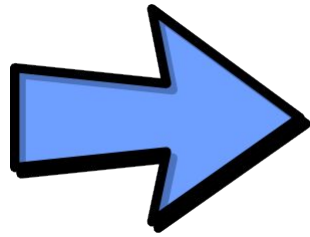
פרק 2 - מאגרי נתונים וטבלאות

# תוכנית המפגש

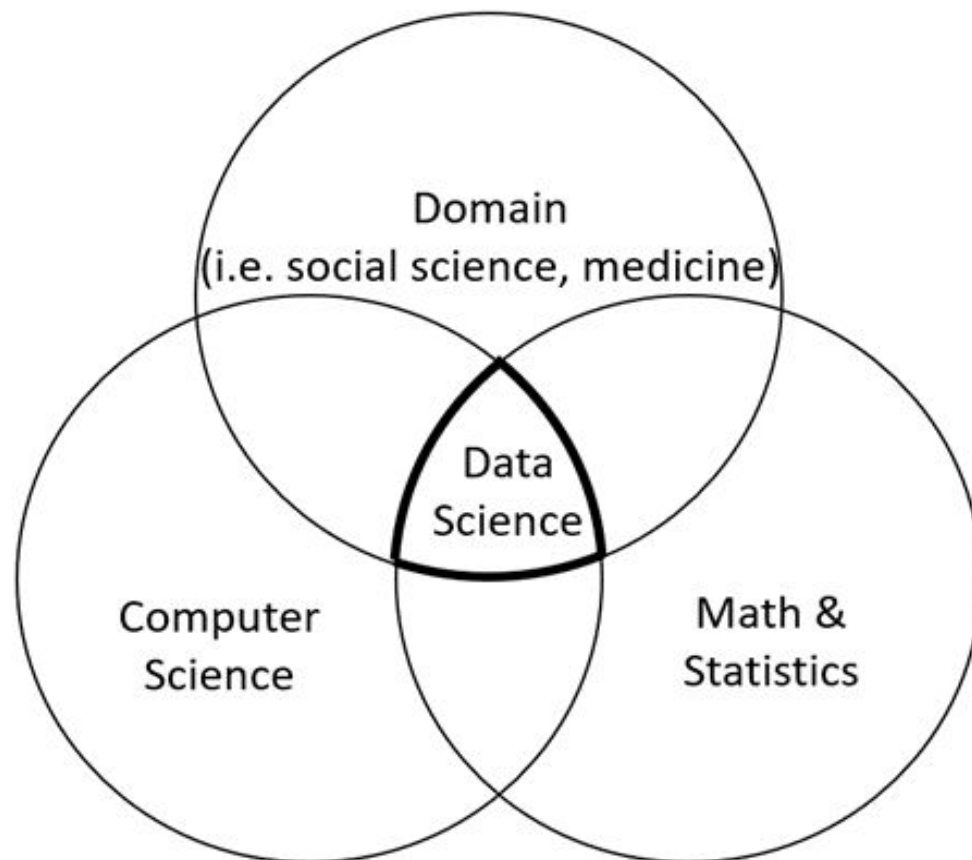
- מבוא למדעי הנתונים
- תהליך העבודה במדעי הנתונים
- תרגול סביבת העבודה
- ספריית Pandas

# מדעי הנתונים

Data science is a new science, focused on generating value from data  
(Skiena, 2017)

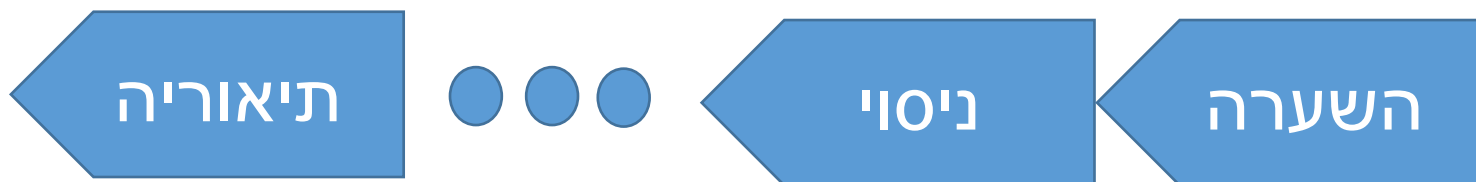


# מדעי הנתונים הוא מדע בין תחומי

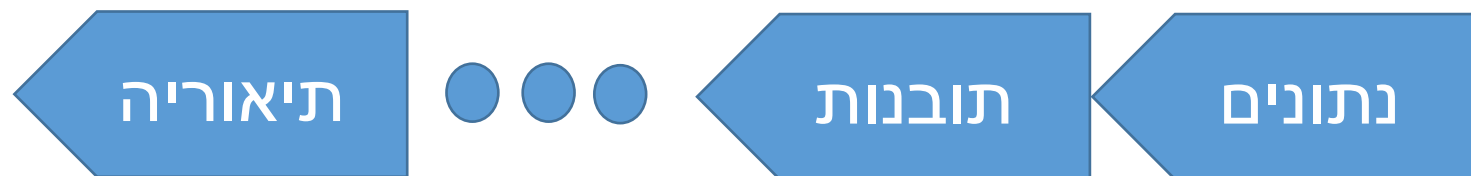


# מדעי הנתונים - פרדיגמה מדעית חדשה

הפרדיגמה המדעית



מדעי הנתונים מאפשרים פרדיגמה חדשה



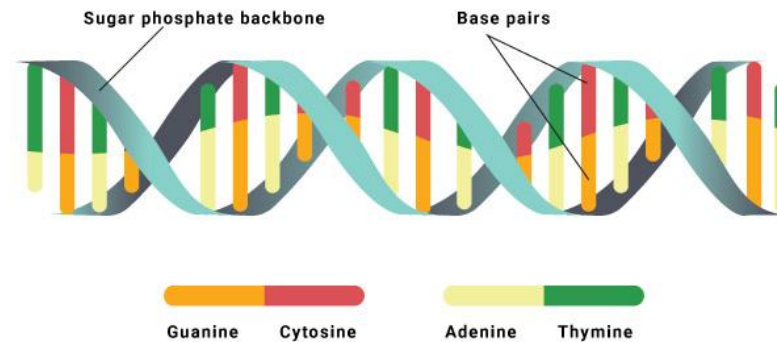
# דוגמאות ליישומי מדעי הנתונים

## Education



<https://www.youtube.com/watch?v=U1eHbHHhI>

## Biomedical



[https://www.youtube.com/watch?v=t7bNe\\_Y2Pag](https://www.youtube.com/watch?v=t7bNe_Y2Pag)

## Social data



[https://www.youtube.com/watch?v=JAO\\_3EvD3DY](https://www.youtube.com/watch?v=JAO_3EvD3DY)

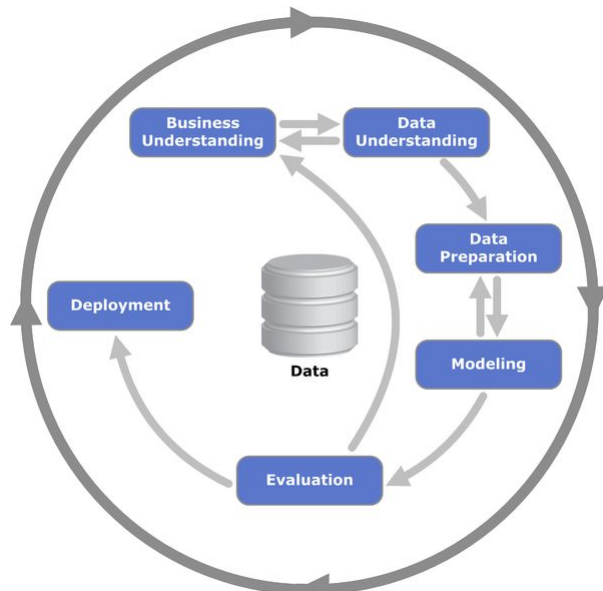


**הטכניון**  
מכון טכנולוגי  
לישראל

**תהליך העבודה במדעי  
הנתונים**

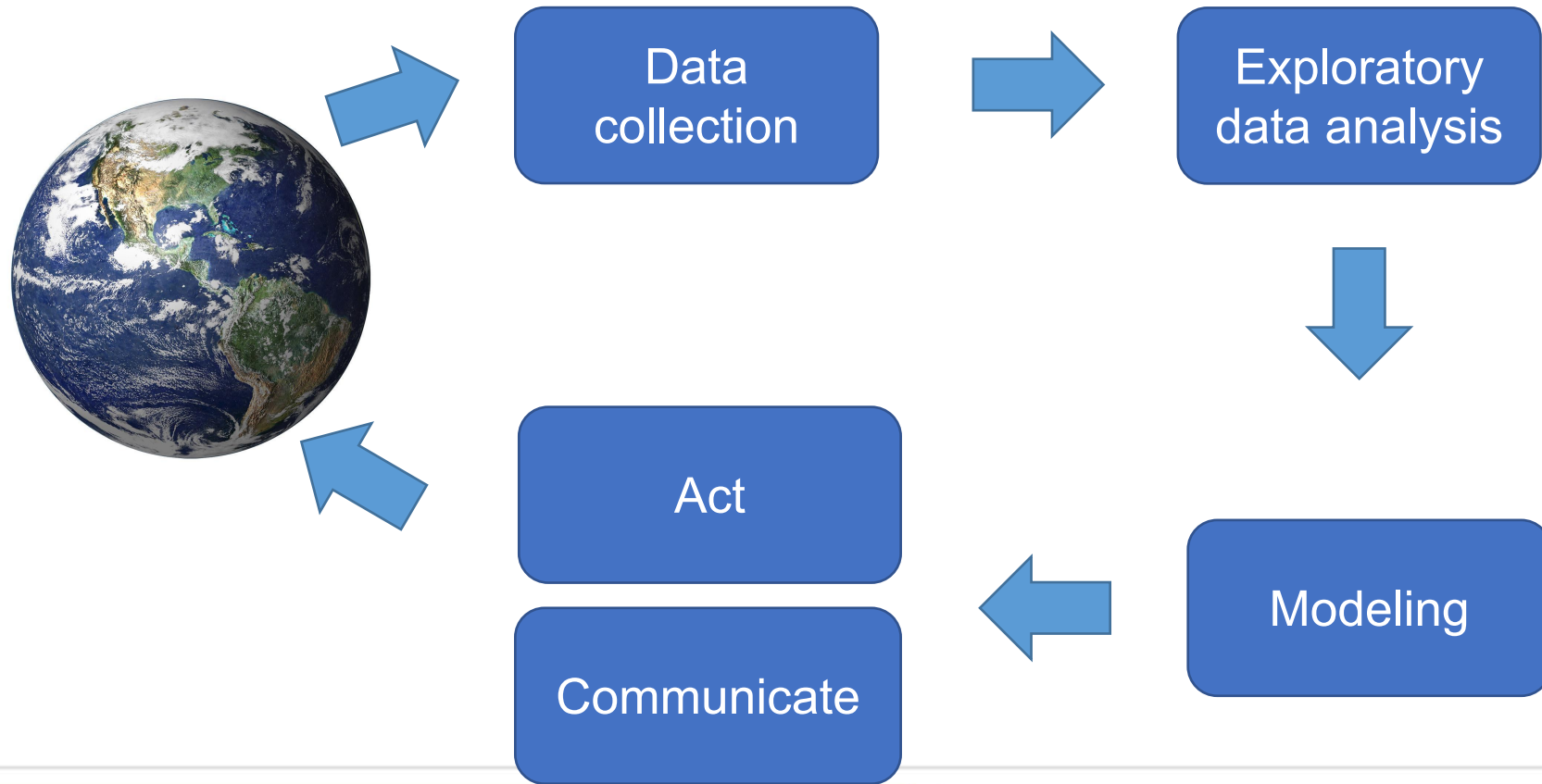
# תהליך העבודה במדעי הנתונים

- אין תהליך יחיד המוסכם באופן רחב באקדמיה ובתעשייה.
- יש מספר תהליכים נפוצים.
- מאפיינים רבים משותפים בין תהליכים אלו.

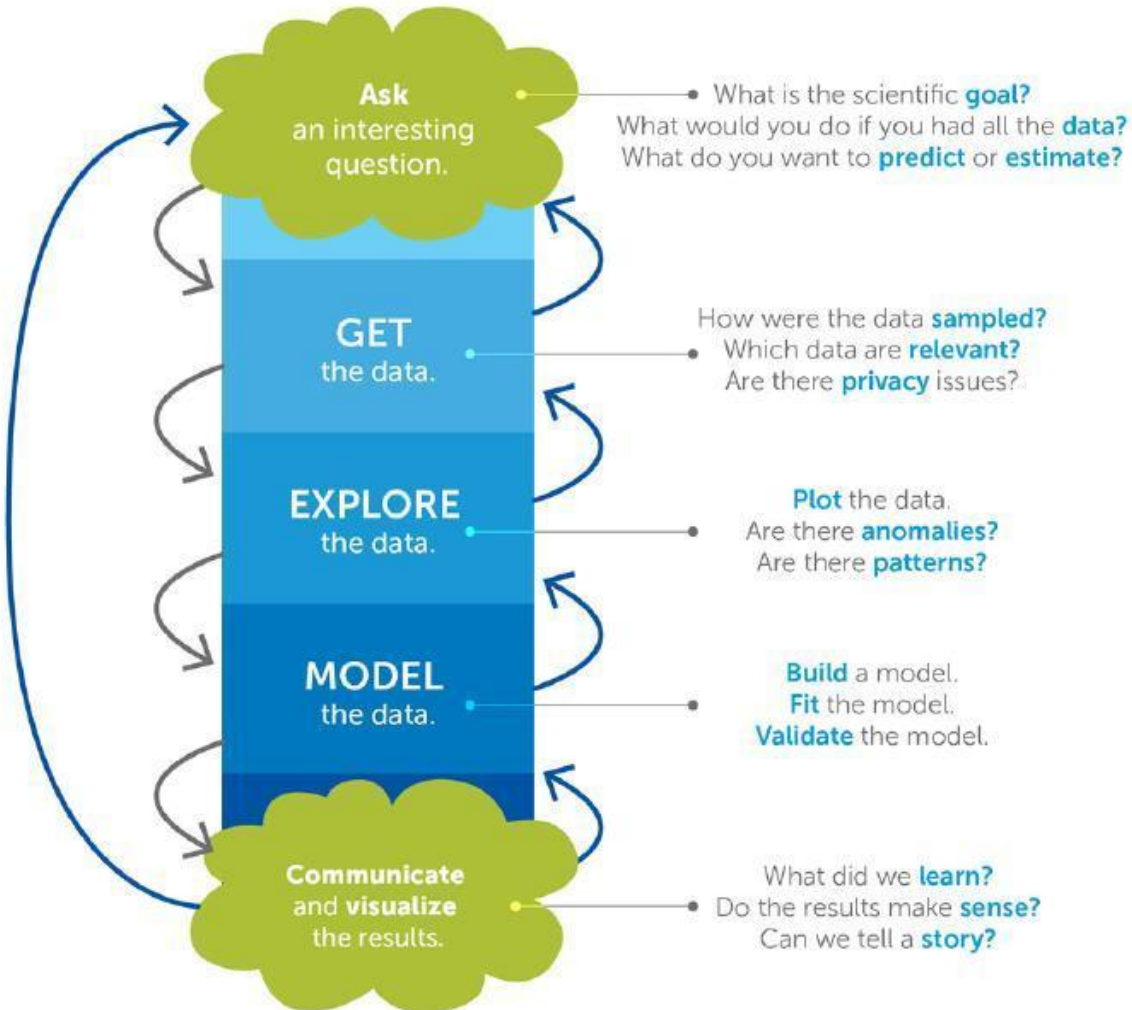


[https://commons.wikimedia.org/wiki/File:CRISP-DM\\_Process\\_Diagram.png](https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)

# תהליך העבודה במדעי הנתונים (2)



# תהליך אג'ילי למדעי הנתונים



**i** Derived by Opera Solutions from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course <http://cs109.org/>.

<https://academy.vertabelo.com/blog/agile-data-science-improve-workflow-with-scrum/>

# דוגמא למחקר במדעי הנתונים: TED



# שלב 1 - שאלות מחקר

- מה מעניין אותנו לדעת על TED?



Ask  
an interesting  
question.

GET  
the data.

EXPLORE  
the data.

MODEL  
the data.

Communicate  
and visualize  
the results.

# שלב 2 - איסוף נתונים

- אילו נתונים נדרשים?
- איך נאסוף את הנתונים הנדרשים?



Ask  
an interesting  
question.

GET  
the data.

EXPLORE  
the data.

MODEL  
the data.

Communicate  
and visualize  
the results.

# איסוף נתונים גולמיים

Web scraping •

The screenshot shows the TED website's 'TED Recommends' section. At the top, the TED logo and 'Ideas worth spreading' are visible. The navigation menu includes 'WATCH', 'DISCOVER', 'ATTEND', 'PARTICIPATE', 'ABOUT', and 'LOG IN'. The main content area features a large background image of two speakers. The text 'TED Recommends' is prominently displayed, followed by 'Talks recommended just for you, delivered to your inbox'. Below this, a question 'What interests you?' is followed by a grid of 15 interest tags: Technology, Science, Design, Business, Collaboration, Innovation, Social change, Health, Nature, The environment, The future, Communication, Activism, Child development, Personal growth, Humanity, Society, Identity, and Community. A 'Next' button is located below the tags. At the bottom of the main section, it says '1/2' and 'Already have a TED account? [Sign in](#) to see your recommendations'. Below the main section is a 'Newest Talks' section with a row of video thumbnails. The browser's address bar shows 'https://www.ted.com/#/' and the taskbar at the bottom displays 'nbgrader - nbgr...html' and 'Weka 3 - Data Mi...html'.

# חיפוש נתונים ברשת



ted dataset



All

Images

Videos

News

Maps

More

Settings

Tools

About 4,930,000 results (0.38 seconds)

## TED Talks | Kaggle

<https://www.kaggle.com/rounakbanik/ted-talks>

Sep 25, 2017 - The **TED** main **dataset** contains information about all talks including number of views, number of comments, descriptions, speakers and titles.

## TED Data Analysis | Kaggle

<https://www.kaggle.com/rounakbanik/ted-data-analysis>

The main **dataset** contains metadata about every **TED** Talk hosted on the **TED.com** website until September 21, 2017. Let me give you a brief walkthrough of the ...

## TED Talks- Complete List - dataset by owentemple | data.world

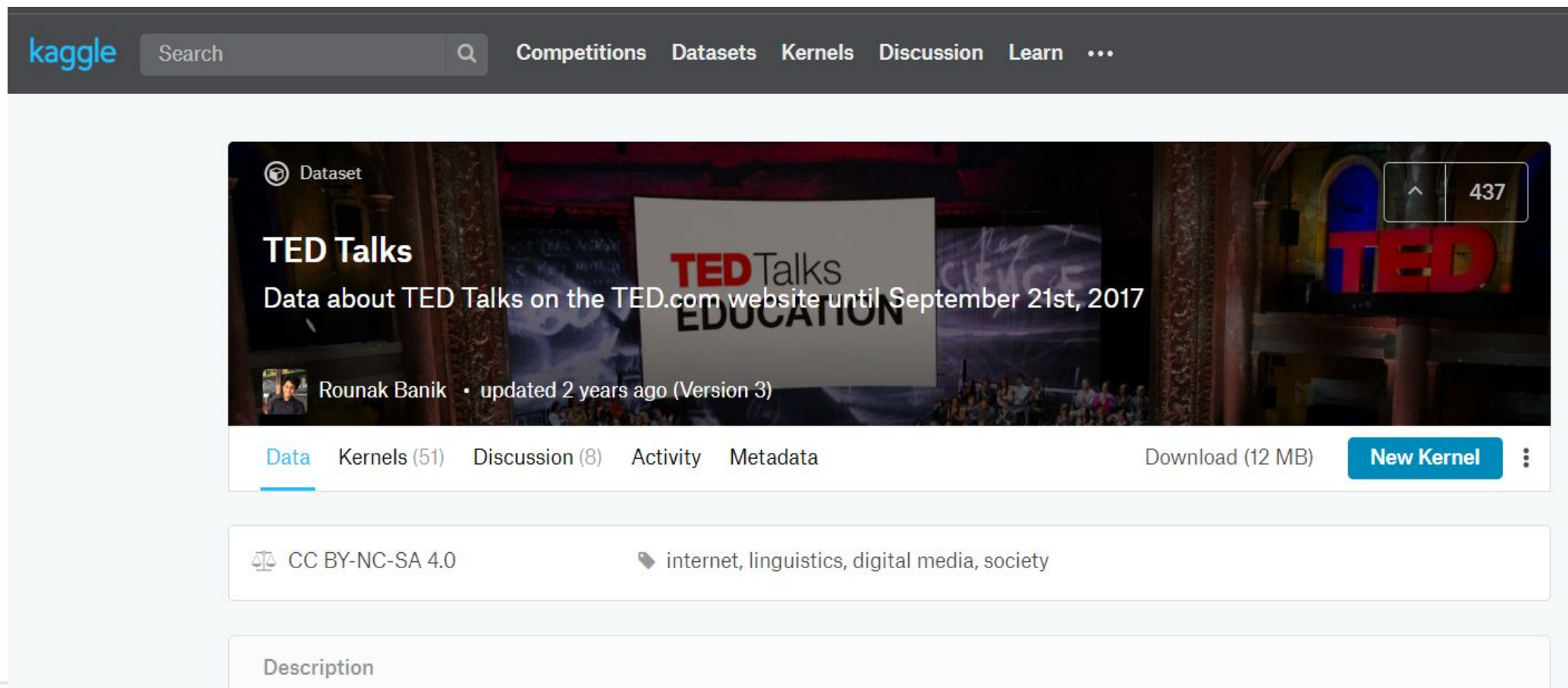
<https://data.world/owentemple/ted-talks-complete-list>

Complete listing of all official **TED** Talks posted (through June 13, 2017) ... An expanded version of the **dataset** includes 111 additional columns with full English ...

Tags: education, technology, entertainment, d... Version: 109994d0



# נתוני TED באתר Kaggle



The screenshot shows the Kaggle website interface. At the top, there is a navigation bar with the Kaggle logo, a search bar, and links for Competitions, Datasets, Kernels, Discussion, and Learn. The main content area features a dataset card for "TED Talks". The card includes a "Dataset" icon, the title "TED Talks", a subtitle "Data about TED Talks on the TED.com website until September 21st, 2017", and the creator's name "Rounak Banik" with a note "updated 2 years ago (Version 3)". Below the card, there are tabs for "Data", "Kernels (51)", "Discussion (8)", "Activity", and "Metadata". To the right of these tabs, there is a "Download (12 MB)" link and a "New Kernel" button. At the bottom of the card, there is a license section showing "CC BY-NC-SA 4.0" and a list of tags: "internet, linguistics, digital media, society". A "Description" section is partially visible at the bottom of the card.

# שלב 3 - חקר נתונים



- המטרה - הבנת תופעות בנתונים לטובת:
  - מיקוד שאלות המחקר
  - זיהוי ממצאים מעניינים
  - זיהוי טעויות, חריגים, נתונים חסרים
- השיטה
  - חישוב מדדים סטטיסטיים
  - ויזואליזציה

ספריית Pandas  
ניהול טבלאות

ספריית Seaborn  
ויזואליזציה

# שלב 3 - חקר נתונים - דוגמא TED

נוכל לבצע חקר כפי שעשינו בשיעורים הקודמים:  
נאתר את בסיס הנתונים של TED באתר Kaggle  
נוריד את הנתונים למחשב  
נטען את הנתונים לתוכנת Excel  
נחקור את הנתונים

# שלב 4 - בניית מודל



- לטובת:
  - יצירת תובנות
  - חיזוי
  - קבלת החלטות
- על ידי:
  - מודלים סטטיסטיים
  - מודלים של למידת מכונה



# שלב 5 - תוצר מחקרי



- לספר את הסיפור של הנתונים
- להמחיש את הסיפור של הנתונים
- לייצר תגובה בעולם הפיזי
- ביחידה שלנו – התוצר הוא הפרויקט



**הטכניון**  
מכון טכנולוגי  
לישראל

# Python notebooks

# Google Colab

- סביבת הרצה לתוכניות פיתון בתוך דפדפן.
- גישה נוחה לקבצים המאוחסנים ב-google drive.
- חסרונות:
  - דורשת קישוריות לרשת
  - מהירות הביצוע לעיתים איטית

# סביבת עבודה מסוג notebooks

- הקוד ותוצאות הרצתו מופיעים יחד באותו החלון זה אחר זה

```
[1] print ('Python')
```

```
↳ Python
```

```
[2] print ('is')
```

```
↳ is
```

```
[3] print ('fun ')
```

```
↳ fun
```

# סביבת העבודה סרטון הסבר



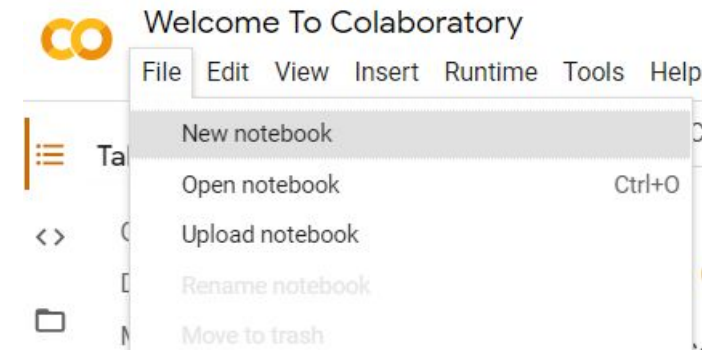
הטכניון  
מכון טכנולוגי  
לישראל

## Python - intro

© kobymike@gmail.com



# הדגמה - סביבת עבודה



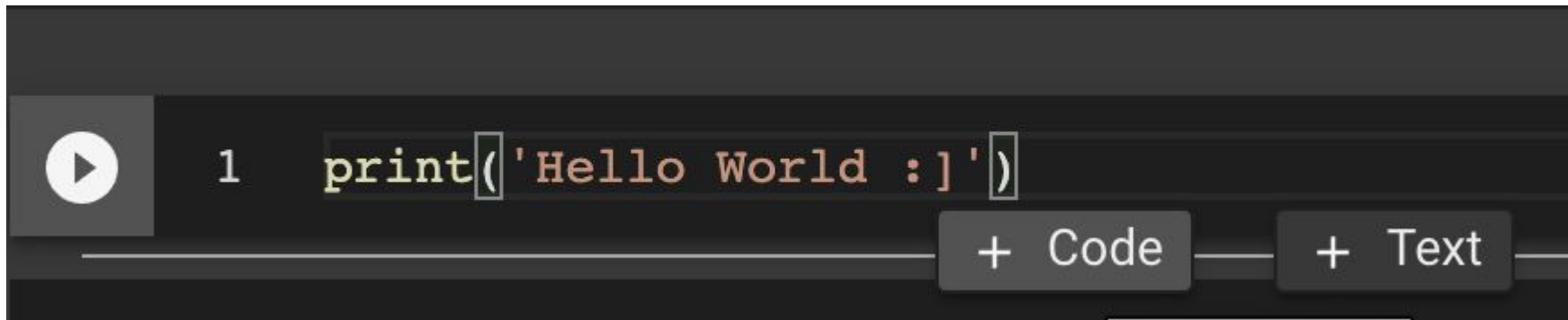
- הפעילו את [google colab](https://colab.research.google.com/).
- פתחו מחברת חדשה.
- הקלידו:

```
print('Hello World!')
```

- הריצו ע"י לחיצה על לחצן ההרצה.

# הדגמה - סביבת עבודה

- הוסיפו תא מתחת או מעל לתא הנוכחי:  
מקמו את העבר על הגבול התחתון או העליון של התא, עד שיופיעו הכפתורים. לחצו על הכפתור Code.
- הדפיסו את שמכם בתא החדש שיצרתם.



```
1 print('Hello World :|')
```

+ Code + Text



**הטכניון**  
מכון טכנולוגי  
לישראל

# ספריית Pandas

# ספריית Pandas

- ספרייה לניהול טבלאות (data frames)

# ייבוא הספרייה לסביבת העבודה



```
import pandas as pd
```



# קבצים: סרטון הסבר

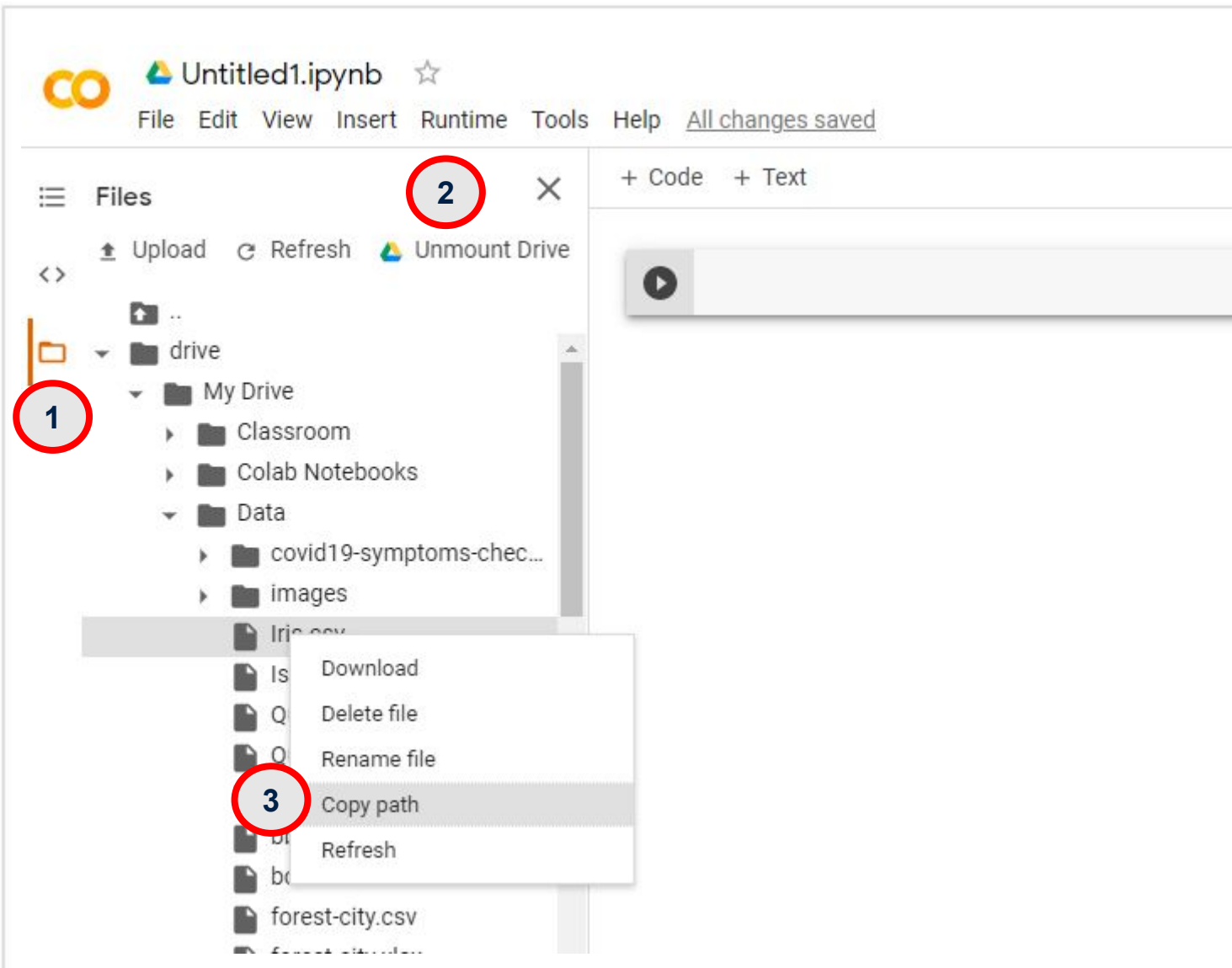


הטכניון  
מכון טכנולוגי  
לישראל

קלט מקבצים

© kobvmike@gmail.com



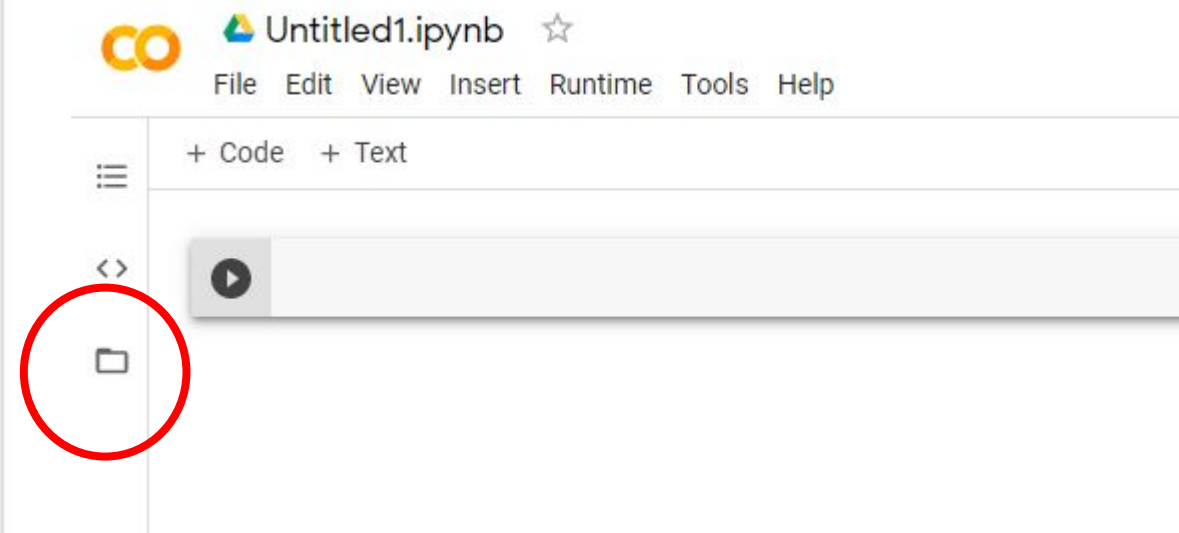


# טעינת קובץ

1. הורידו למחשב שלכם את הקובץ [בלינק הזה](#).
2. העלו את הקובץ לדרייב של הקלאסרום.  
הדרייב נמצא בלשונית עבודת כיתה בקלאסרום.
3. חזרו למחברת קולאב:
  - 3.1. הקליקו על האייקון של הקבצים.
  - 3.2. הקליקו על האייקון של גוגל דרייב, ואשרו שימוש בדרייב שלכם.
  - 3.3. לאחר חיבור הדרייב, תוכלו להשתמש במחברת בקבצים שבדרייב בעזרת העתקת כתובת הקובץ.

# חיבור סביבת העבודה ל-google drive

נרצה לקשר בין המחברת לדרייב כדי שנוכל להשתמש בקבצי נתונים מהדרייב שלנו.  
ראשית, לחצו על האייקון של הקבצים בצד שמאל של המחברת.



# דוגמא: קריאת טבלה מקובץ CSV

1

```
df = pd.read_csv('/content/drive/My Drive/Data/datasets_19_420_Iris.csv')  
df
```

2

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

1. הקלידו את הפקודה לטעינת הנתונים,

ובשורה הבאה את הפקדה להצגה כללית של הטבלה.

השתמשו בכתובת הקובץ שהעתקתם במסך הקודם. ❤️

2. הריצו את התא שבמחברת.

3. הקובץ מכיל נתונים על סוגים שונים של פרחי אירוס.



# תכונות בסיסיות של טבלאות pandas

- הצגת ראש הטבלה
- בחירת שמות העמודות
- בחירת עמודות
- בחירת שורות
- מיון
- חישוב מדדים סטטיסטיים (סטטיסטיקה תיאורית)

# Pandas

## סרטון הסבר



הטכניון  
מכון טכנולוגי  
לישראל

## ספריית Pandas

@kotbymike@gmail.com



# הצגת ראש הטבלה – head

```
df.head()
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa



# הצגת שמות העמודות – columns

```
df.columns
```

```
Index(['sepal.length', 'sepal.width', 'petal.length', 'petal.width',  
      'variety'],  
      dtype='object')
```

# בחירת עמודה - col

אפשרות ב': התוצאה היא טבלה

```
[ ] df[['sepal.length']]
```

	sepal.length
0	5.1
1	4.9
2	4.7
3	4.6
4	5.0
...	...
145	6.7
146	6.3
147	6.5
148	6.2
149	5.9

150 rows × 1 columns

אפשרות א': התוצאה היא רשימה

```
df['sepal.length']
```

0	5.1
1	4.9
2	4.7
3	4.6
4	5.0
...	...
145	6.7
146	6.3
147	6.5
148	6.2
149	5.9

Name: sepal.length, Length: 150, dtype: float64

# בחירת עמודות

```
df[['sepal.length', 'sepal.width']]
```

	sepal.length	sepal.width
0	5.1	3.5
1	4.9	3.0
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6
...	...	...
145	6.7	3.0
146	6.3	2.5
147	6.5	3.0
148	6.2	3.4
149	5.9	3.0

150 rows × 2 columns



# בחירת שורות לפי אינדקס from:to

```
df[0:10]
```

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
5	5.4	3.9	1.7	0.4	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa



# בחירת שורות באמצעות תנאי

```
df[df['variety']=='Virginica']
```

	sepal.length	sepal.width	petal.length	petal.width	variety
100	6.3	3.3	6.0	2.5	Virginica
101	5.8	2.7	5.1	1.9	Virginica
102	7.1	3.0	5.9	2.1	Virginica
103	6.3	2.9	5.6	1.8	Virginica
104	6.5	3.0	5.8	2.2	Virginica
105	7.6	3.0	6.6	2.1	Virginica
106	4.9	2.5	4.5	1.7	Virginica
107	7.3	2.9	6.3	1.8	Virginica
108	6.7	2.5	5.8	1.8	Virginica
109	7.2	3.6	6.1	2.5	Virginica



# בחירת שורות באמצעות תנאי (2)

```
df[df['petal.length'] > 6]
```

	sepal.length	sepal.width	petal.length	petal.width	variety
105	7.6	3.0	6.6	2.1	Virginica
107	7.3	2.9	6.3	1.8	Virginica
109	7.2	3.6	6.1	2.5	Virginica
117	7.7	3.8	6.7	2.2	Virginica
118	7.7	2.6	6.9	2.3	Virginica
122	7.7	2.8	6.7	2.0	Virginica
130	7.4	2.8	6.1	1.9	Virginica
131	7.9	3.8	6.4	2.0	Virginica
135	7.7	3.0	6.1	2.3	Virginica



# בחירת שורות באמצעות תנאי מורכב

```
df[(df['variety'] == 'Virginica') & (df['petal.length'] > 6)]
```

	sepal.length	sepal.width	petal.length	petal.width	variety
105	7.6	3.0	6.6	2.1	Virginica
107	7.3	2.9	6.3	1.8	Virginica
109	7.2	3.6	6.1	2.5	Virginica
117	7.7	3.8	6.7	2.2	Virginica
118	7.7	2.6	6.9	2.3	Virginica
122	7.7	2.8	6.7	2.0	Virginica
130	7.4	2.8	6.1	1.9	Virginica
131	7.9	3.8	6.4	2.0	Virginica
135	7.7	3.0	6.1	2.3	Virginica



# מיון

```
df.sort_values(by='petal.length')
```

	sepal.length	sepal.width	petal.length	petal.width	variety
22	4.6	3.6	1.0	0.2	Setosa
13	4.3	3.0	1.1	0.1	Setosa
14	5.8	4.0	1.2	0.2	Setosa
35	5.0	3.2	1.2	0.2	Setosa
36	5.5	3.5	1.3	0.2	Setosa
...	...	...	...	...	...
131	7.9	3.8	6.4	2.0	Virginica
105	7.6	3.0	6.6	2.1	Virginica
117	7.7	3.8	6.7	2.2	Virginica
122	7.7	2.8	6.7	2.0	Virginica
118	7.7	2.6	6.9	2.3	Virginica

150 rows × 5 columns



# ביצוע חישובים

```
df['petal.length.in'] = df['petal.length'] / 2.54  
df
```

	sepal.length	sepal.width	petal.length	petal.width	variety	petal.length.in
0	5.1	3.5	1.4	0.2	Setosa	0.551181
1	4.9	3.0	1.4	0.2	Setosa	0.551181
2	4.7	3.2	1.3	0.2	Setosa	0.511811
3	4.6	3.1	1.5	0.2	Setosa	0.590551
4	5.0	3.6	1.4	0.2	Setosa	0.551181
...	...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Virginica	2.047244
146	6.3	2.5	5.0	1.9	Virginica	1.968504
147	6.5	3.0	5.2	2.0	Virginica	2.047244
148	6.2	3.4	5.4	2.3	Virginica	2.125984
149	5.9	3.0	5.1	1.8	Virginica	2.007874

150 rows × 6 columns



# סטטיסטיקה תיאורית

```
df.describe()
```

	sepal.length	sepal.width	petal.length	petal.width
<b>count</b>	150.000000	150.000000	150.000000	150.000000
<b>mean</b>	5.843333	3.057333	3.758000	1.199333
<b>std</b>	0.828066	0.435866	1.765298	0.762238
<b>min</b>	4.300000	2.000000	1.000000	0.100000
<b>25%</b>	5.100000	2.800000	1.600000	0.300000
<b>50%</b>	5.800000	3.000000	4.350000	1.300000
<b>75%</b>	6.400000	3.300000	5.100000	1.800000
<b>max</b>	7.900000	4.400000	6.900000	2.500000



# תרגיל

- טענו את [טבלת נתוני TED](#) לסביבת colab באמצעות Pandas ובצעו את המשימות למטה.
- ❤️ עליכם להגיש קובץ עם תשובות לשאלות המידע הסטטיסטי, ואת המחברת שיצרתם - עם **הרשאת צפייה** פתוחה לכל העולם.

- הצגת נתונים:

- הציגו טבלה הכוללת את העמודות 'title' ו 'views' בלבד
  - בחרו את כל ההרצאות מהאירוע TED2006
  - בחרו את כל ההרצאות שיש להן מעל 10,000,000 צפיות
  - בחרו את כל ההרצאות מהאירוע TED2006 שיש להן מעל 10,000,000 צפיות
- הערה: פתרו כל סעיף בנפרד – אין קשר בין הסעיפים

- מידע סטטיסטי:

- מהו ממוצע הצפיות של כלל ההרצאות בקובץ?
- מהו ממוצע הצפיות של ההרצאות בTED2006?
- מהו ממוצע אורך ההרצאות של כלל ההרצאות בקובץ?
- מהו ממוצע אורך ההרצאות בTED2006?



**הטכניון**  
מכון טכנולוגי  
לישראל

**תודה על ההשתתפות**