University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Data analysis and visualization (DAV)

Lecture 03

Łukasz P. Kozłowski

l.kozlowski@mimuw.edu.pl

Warsaw, 2025

University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Data analysis and visualization (DAV)

Lecture 03

Basic tools & datasets

Łukasz P. Kozłowski

Warsaw, 2025

l.kozlowski@mimuw.edu.pl

Python was conceived in the late 1980s

Guido van Rossum

Guido van Rossum

Python was conceived in the late 1980s

https://en.wikiversity.org/wiki/Python_Concepts

https://en.wikiversity.org/wiki/Python_Programming

https://pl.wikibooks.org/wiki/Zanurkuj_w_Pythonie

Coursera LinkedIn Learning Datacamp, Inc. edX Udacity

Guido van Rossum

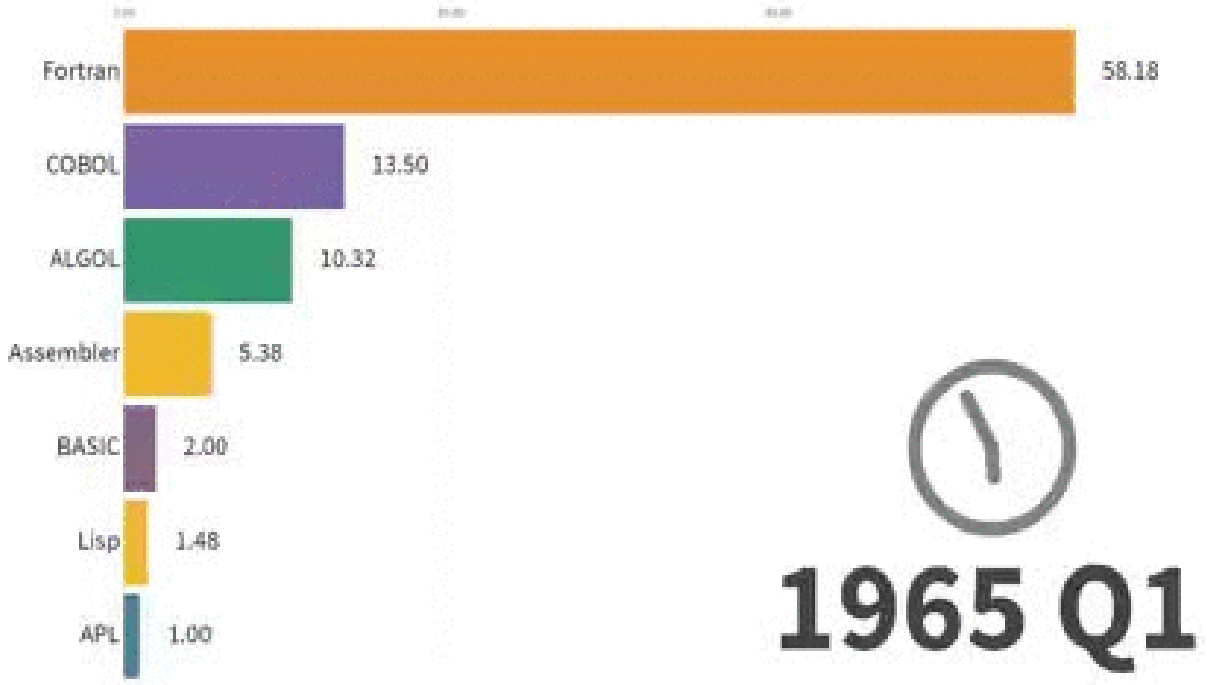
Python was conceived in the late 1980s

- scripting language (no compilation)

- uses whitespace indentation, rather

than curly brackets or keywords,

to delimit blocks



- we use **python3** instead **python2**

https://www.youtube.com/watch?v=Og847HVwRSI

PYTHON 2.X

LEGACY

It Is stlll entrenched in the software at certain companles

LIBRARY

Many older libraries built for Python 2 are not forwards compatible

ASCII

Strings are stored as ASCII by default

7/2=3

It rounds your calculation down to the nearest whole number

print "WELCOME TO GEEKSFORGEEKS"

It rounds your calculation down to the nearest whole number

PYTHON 3.X

FUTURE

It will take over Python 2 by the end of 2019

LIBRARY

Many of today's developers are creating libraries strictly for use with Python 3

UNICODE

Text Strings are Unicode by default

7/2=3.5

This expression will result in the expected result

print("WELCOME TO GEEKSFORGEEKS")

This expression will result in the expected result

Range Function

Python 2.7 :

It has both `range` and `xrange` function. When you need to iterate one object at a time, use xrange and when you need an actual list, use range function. xrange is generally faster & saves memory.

```
% timeit [i for i in range(1000)]
% timeit [i for i in xrange(1000)]
```

Python 3.x :

Here range does what xrange does in Python 2.7. xrange doesn't work in Python 3.x.

```
% timeit [i for i in range(1000)]
% timeit [i for i in xrange(1000)]
```

List Comprehensions

Python 2.7: Refer to the example below, how global variable changes.

```
num = 7
print (num)

mylist = [num for num in range(100)]
print (num)

# Output:
# 7
# 99
```

Python 3.x: There is no namespace leak now. This is quite fixed now.

```
num = 7
print (num)

mylist = [num for num in range(100)]
print (num)

# Output:
# 7
# 7
```

Exception Handling

Python 2.7 : This has a different syntax than Python 3.x.

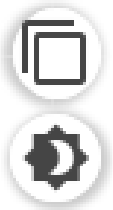
```
try:
    YoYo
except NameError, error:
    print error, "YOU HAVE REACHED FOR AN ERROR"

try:
    YoYo
except NameError as error:
    print error, "YOU HAVE REACHED AN ERROR, YET AGAIN !"
```

Python 3.x: 'As' keyword is needed to be included in this.

```
try:
    YoYo
except NameError as error:
    print (error, "THE ERROR HAS ARRIVED !")
```

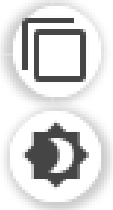
next() function and .next() method

Python 2.7: Both next() and .next() are used here.



```
generator = (letter for letter in 'abcdefg')
next(generator)
generator.next()
```

Python 3.x: Only next() is used here. Using .next() shows an AttributeError.



```
generator = (letter for letter in 'abcdefg')
next(generator)
```

Returning iterable objects instead of lists

Python 2

```
print 'Python', python_version()

print range(3)
print type(range(3))
```

```
Python 2.7.6
[0, 1, 2]
<type 'list'>
```

Python 3

```
print('Python', python_version())

print(range(3))
print(type(range(3)))
print(list(range(3)))
```

```
Python 3.4.1
range(0, 3)
<class 'range'>
[0, 1, 2]
```

Returning iterable objects instead of lists

Python 2

```
print 'Python', python_version()

print range(3)
print type(range(3))
```

```
Python 2.7.6
[0, 1, 2]
<type 'list'>
```

Python 3

```
print('Python', python_version())

print(range(3))
print(type(range(3)))
print(list(range(3)))
```

```
Python 3.4.1
range(0, 3)
<class 'range'>
[0, 1, 2]
```

Some more commonly used functions and methods that don't return lists anymore in Python 3:

zip()	dictionary's .keys() method
map()	dictionary's .values() method
filter()	dictionary's .items() method

Advanced unpacking

- You can already do this:

```
>>> a, b = range(2)
>>> a
0
>>> b
1
```

Advanced unpacking

- You can already do this:

```
>>> a, b = range(2)
>>> a
0
>>> b
1
```

- Now you can do this:

```
>>> a, b, *rest = range(10)
>>> a
0
>>> b
1
>>> rest
[2, 3, 4, 5, 6, 7, 8, 9]
```

Advanced unpacking

- You can already do this:

```
>>> a, b = range(2)
>>> a
0
>>> b
1
```

- Now you can do this:

```
>>> a, b, *rest = range(10)
>>> a
0
>>> b
1
>>> rest
[2, 3, 4, 5, 6, 7, 8, 9]
```

- `*rest` can go anywhere:

```
>>> a, *rest, b = range(10)
>>> a
0
>>> b
9
>>> rest
[1, 2, 3, 4, 5, 6, 7, 8]
```

```
>>> *rest, b = range(10)
>>> rest
[0, 1, 2, 3, 4, 5, 6, 7, 8]
>>> b
9
```

Pathlib

- In Python 2, path handling is verbose

```
import os

directory = "/etc"
filepath = os.path.join(directory, "test_file.txt")

if os.path.exists(filepath):
    stuff
```

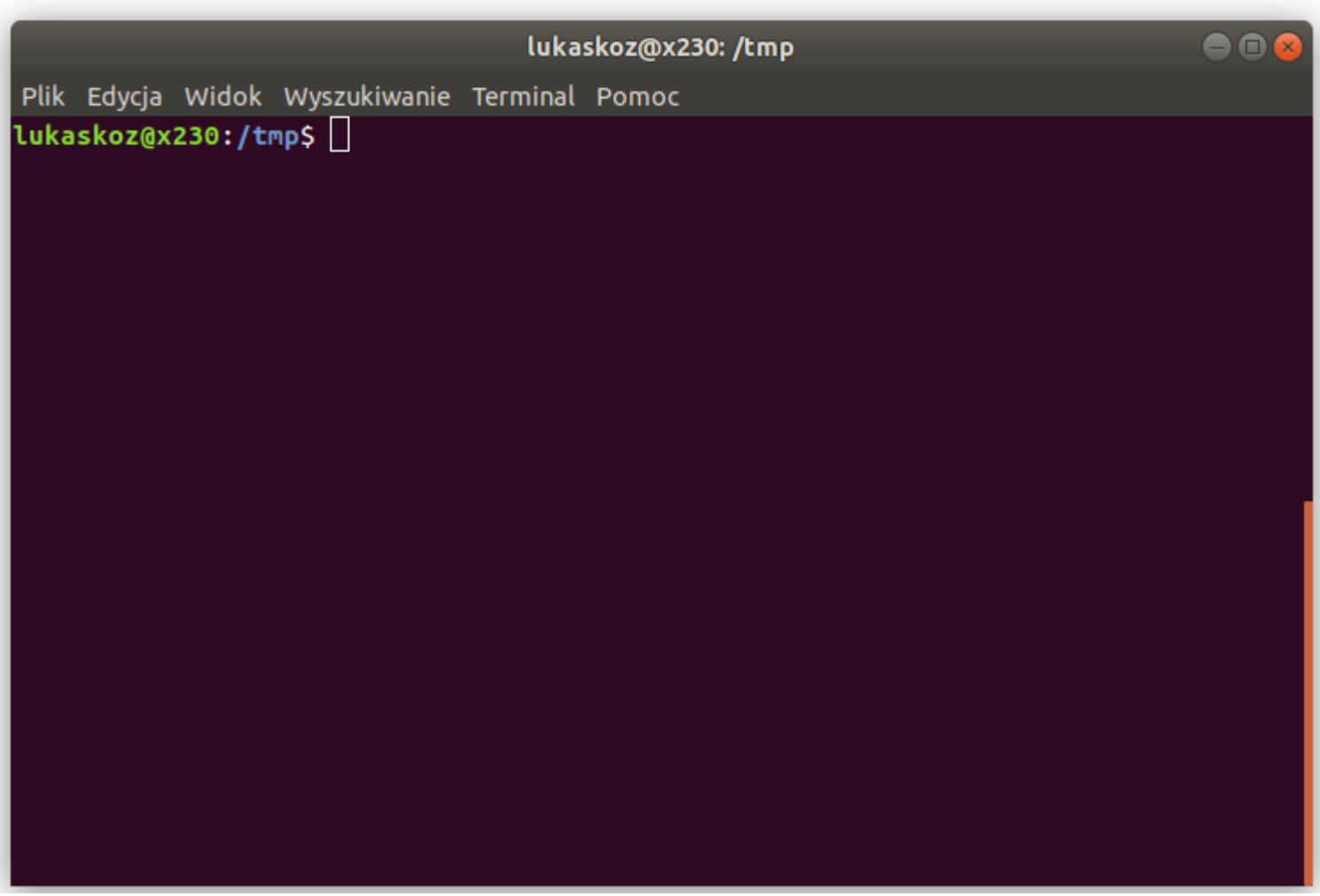
Pathlib

- In Python 2, path handling is verbose

```
import os

directory = "/etc"
filepath = os.path.join(directory, "test_file.txt")

if os.path.exists(filepath):
    stuff
```

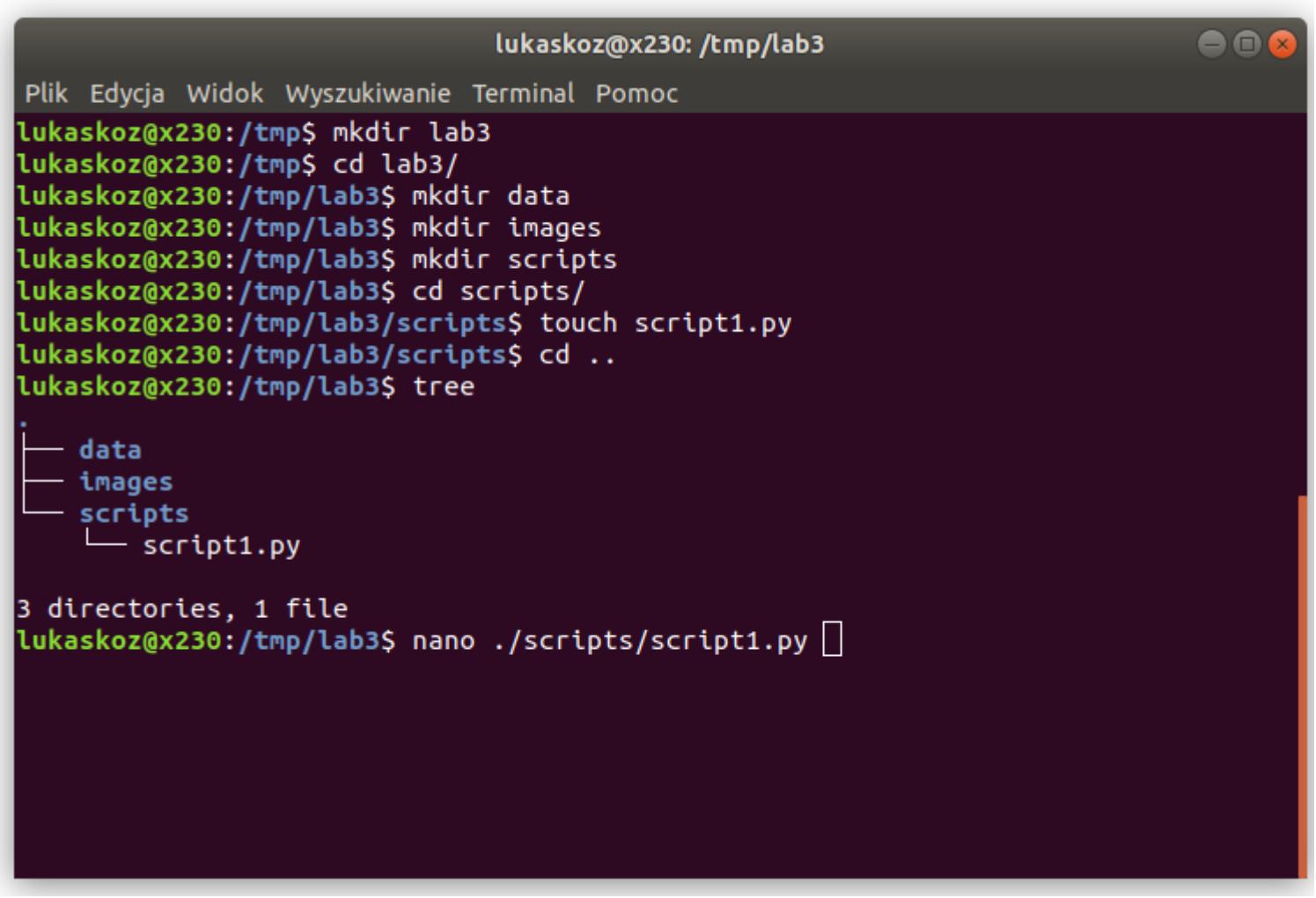
- In Python 3, it is much more simpler

```
from pathlib import Path

directory = Path("/etc")
filepath = directory / "test_file.txt"

if filepath.exists():
    stuff
```



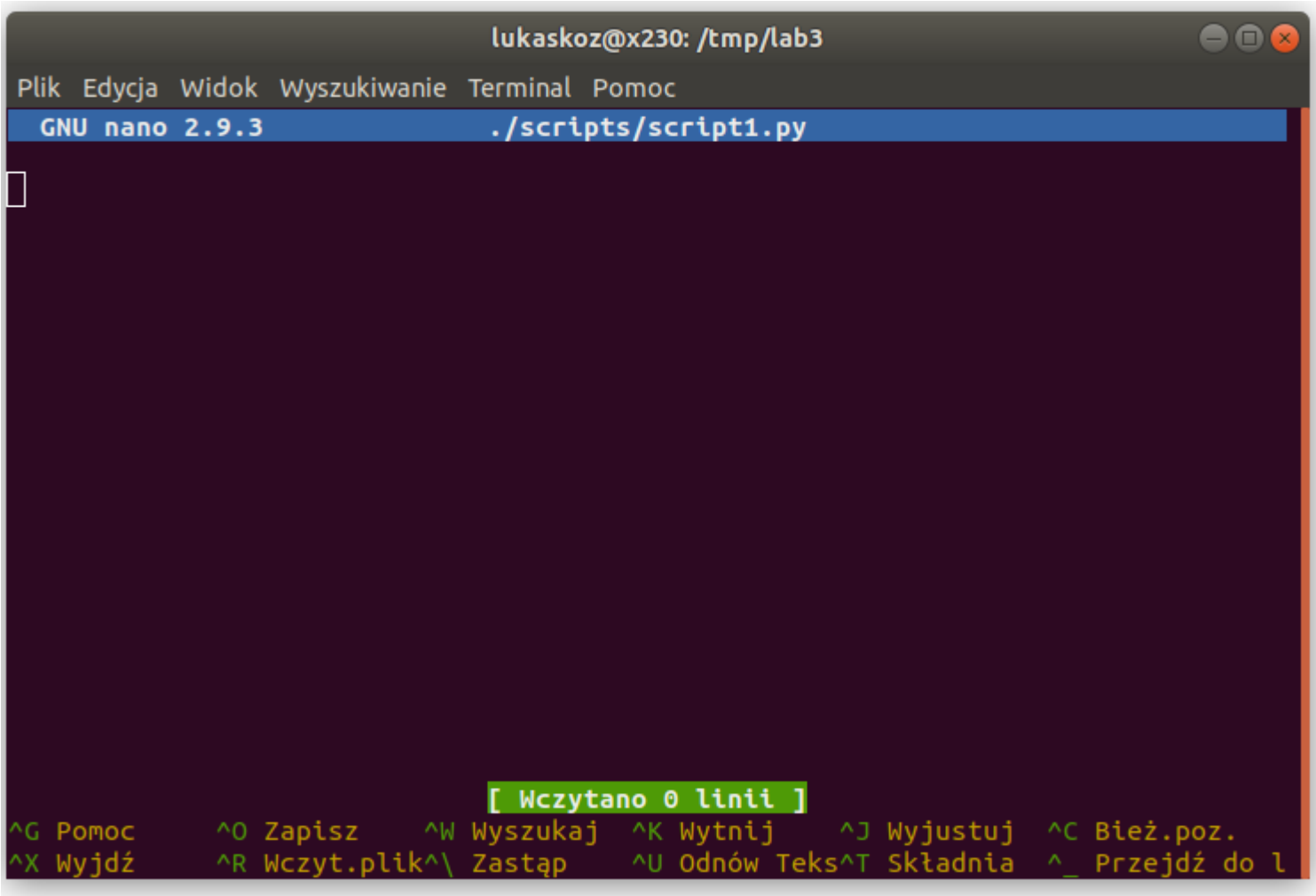
We do as much as possible in PYTHON

The PYTHON works from command line (CLI) very well (Terminal/Console is also installed everywhere)



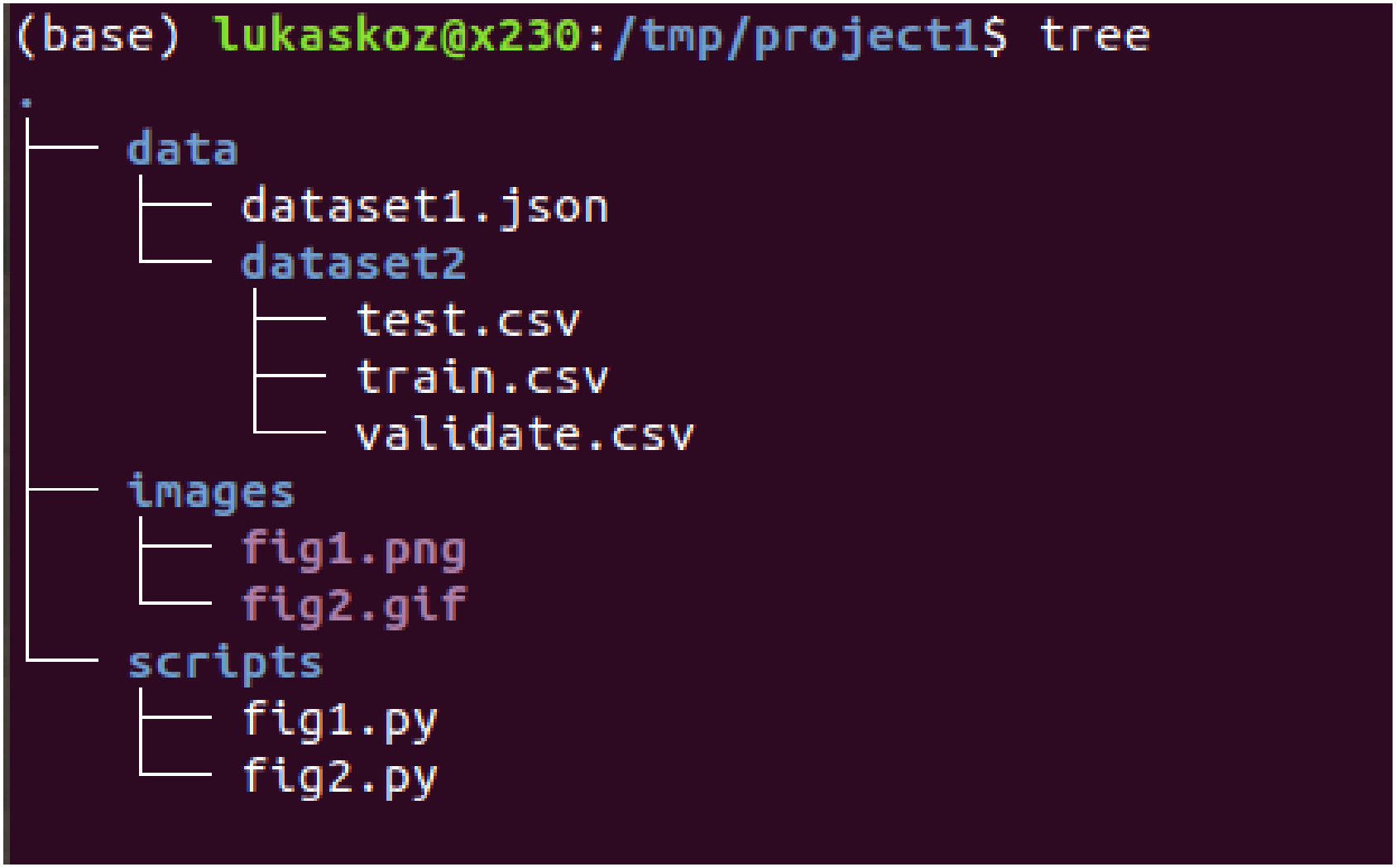
We do as much as possible in PYTHON

The PYTHON works from command line (CLI) very well (Terminal/Console is also installed everywhere)



We do as much as possible in PYTHON

The PYTHON works from command line (CLI) very well (Terminal/Console is also installed everywhere)

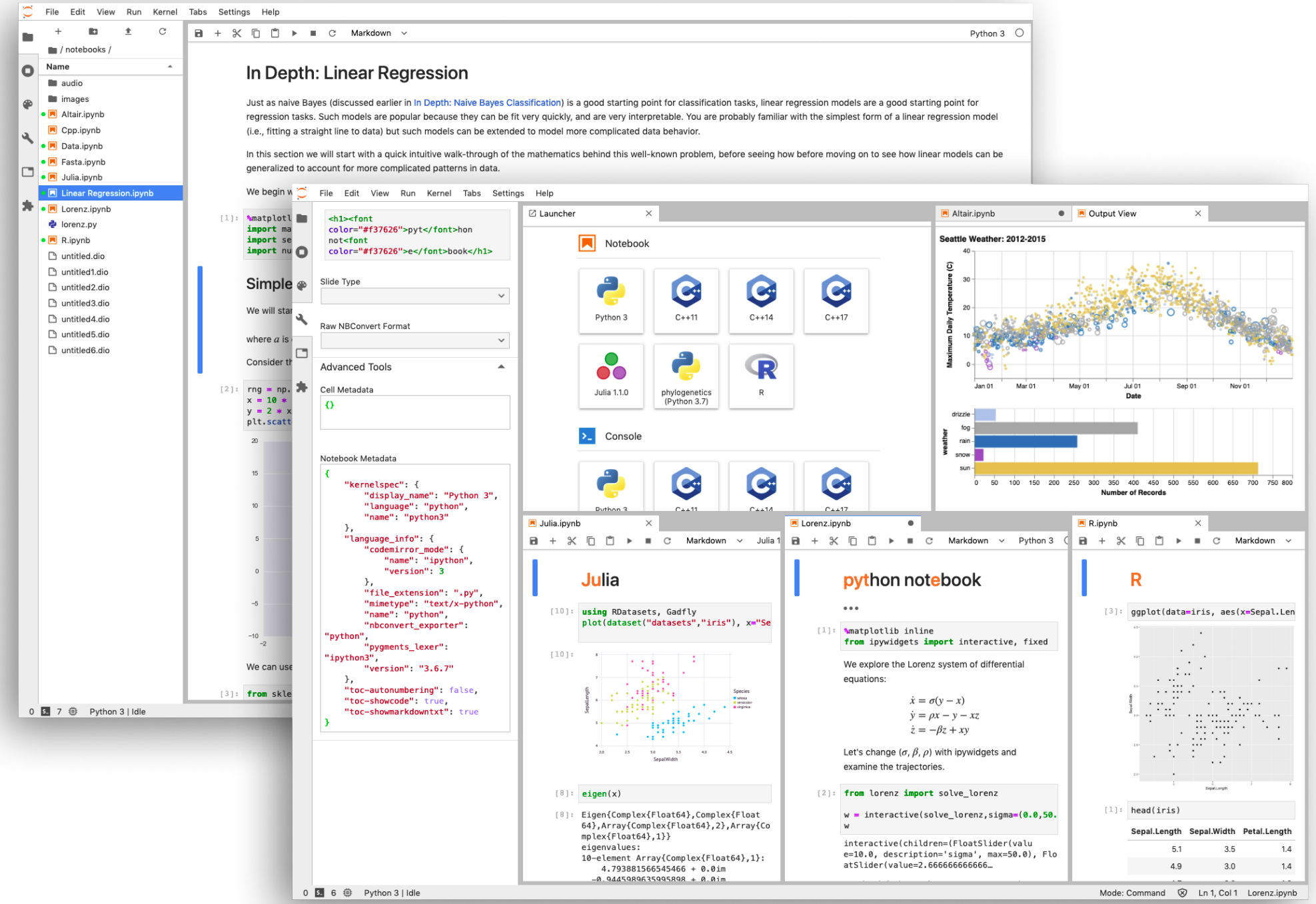

(base) lukaskoz@x230:/tmp/project1$ tree
.
data
dataset1.json
dataset2
test.csv
train.csv
validate.csv
images
fig1.png
fig2.gif
scripts
fig1.py
fig2.py

nano
joe
gedit
kate
mcedit
…

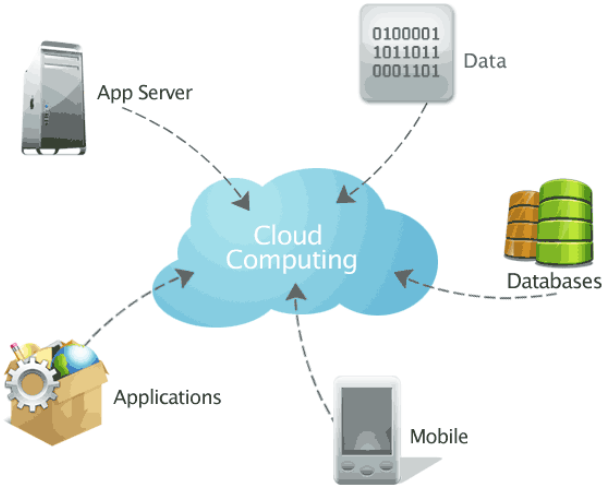
nano
joe
gedit
kate
mcedit
...

Sublime Text

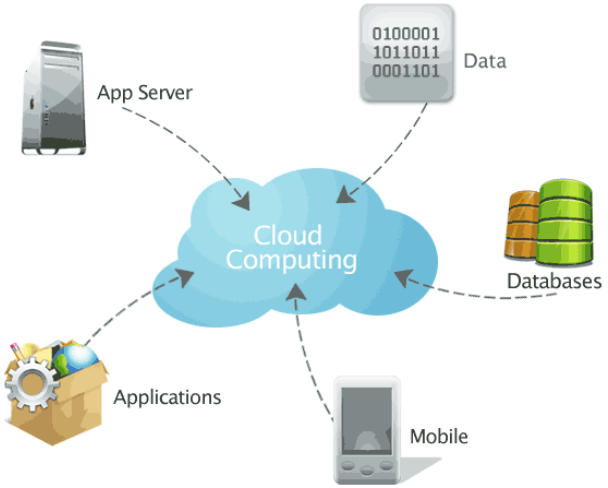

jupyter


File Edit View Run Kernel Tabs Settings Help
Markdown
Python 3
/ notebooks /
Name
audio
images
Altair.ipynb
Cpp.ipynb
Data.ipynb
Fasta.ipynb
Julia.ipynb
Linear Regression.ipynb
Lorenz.ipynb
lorenz.py
R.ipynb
untitled.dio
untitled1.dio
untitled2.dio
untitled3.dio
untitled4.dio
untitled5.dio
untitled6.dio
In Depth: Linear Regression
Just as naive Bayes (discussed earlier in In Depth: Naive Bayes Classification) is a good starting point for classification tasks, linear regression models are a good starting point for regression tasks. Such models are popular because they can be fit very quickly, and are very interpretable. You are probably familiar with the simplest form of a linear regression model (i.e., fitting a straight line to data) but such models can be extended to model more complicated data behavior.
In this section we will start with a quick intuitive walk-through of the mathematics behind this well-known problem, before seeing how before moving on to see how linear models can be generalized to account for more complicated patterns in data.
We begin w
Simple
We will star
where a is
Consider th
We can use
Slide Type
Raw NBConvert Format
Advanced Tools
Cell Metadata
Notebook Metadata
Launcher
Notebook
Python 3
C++11
C++14
C++17
Julia 1.1.0
phylogenetics (Python 3.7)
R
Console
Altair.ipynb
Output View
Seattle Weather: 2012-2015
Maximum Daily Temperature (C)
Date
weather
Number of Records
drizzle
fog
rain
snow
sun
Julia.ipynb
Julia
using RDatasets, Gadfly
plot(dataset("datasets","iris"), x="Se
eigen(x)
Eigen{Complex{Float64},Complex{Float64},Array{Complex{Float64},2},Array{Complex{Float64},1}}
eigenvalues:
10-element Array{Complex{Float64},1}:
4.793881566545466 + 0.0im
Lorenz.ipynb
python notebook
%matplotlib inline
from ipywidgets import interactive, fixed
We explore the Lorenz system of differential equations:
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$
Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.
from lorenz import solve_lorenz
w = interactive(solve_lorenz,sigma=(0.0,50.
interactive(children=(FloatSlider(valu
e=10.0, description='sigma', max=50.0), Flo
atSlider(value=2.666666666666…
R.ipynb
R
ggplot(data=iris, aes(x=Sepal.Len
head(iris)
Sepal.Length Sepal.Width Petal.Length
5.1 3.5 1.4
4.9 3.0 1.4
Python 3 | Idle
Mode: Command
Ln 1, Col 1
Lorenz.ipynb

Work as much as you can on your own (physicial) computer

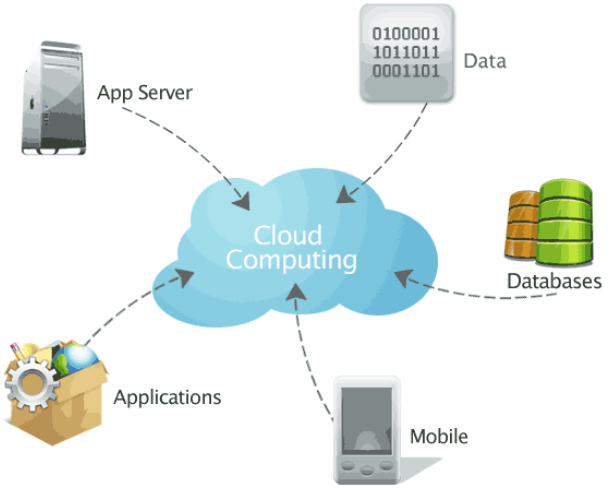


avoid **cloud** or/and **web based solutions** as long as possible



avoid **cloud** or/and **web based solutions** as long as possible

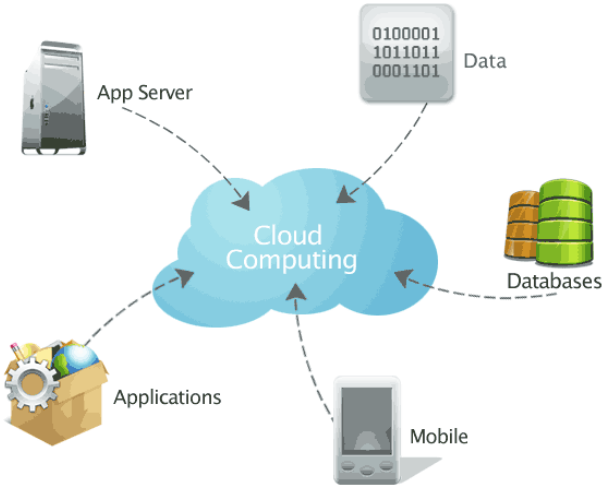
You are loosing the control
(no root, no control on data and resources)



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

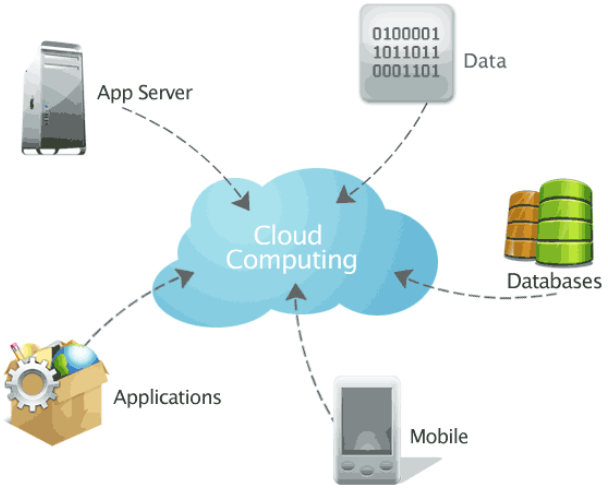


avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up



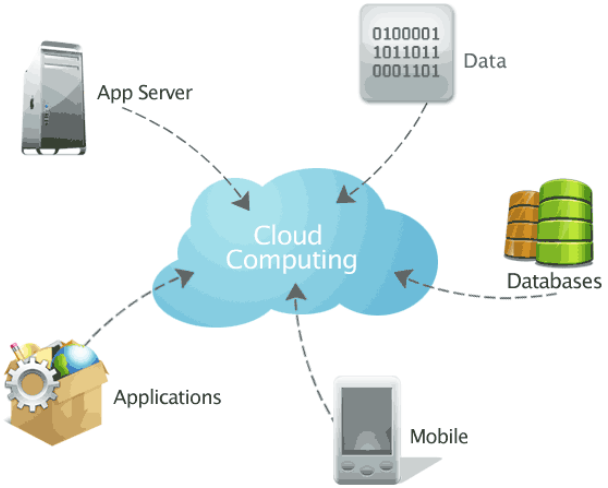
avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

- dependency on service providers



avoid **cloud** or/and **web based solutions** as long as possible

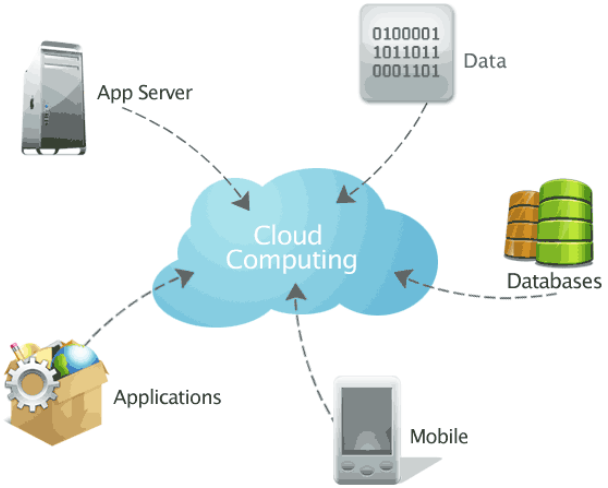
You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

- dependency on service providers

- lack of knowledge and expertise



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

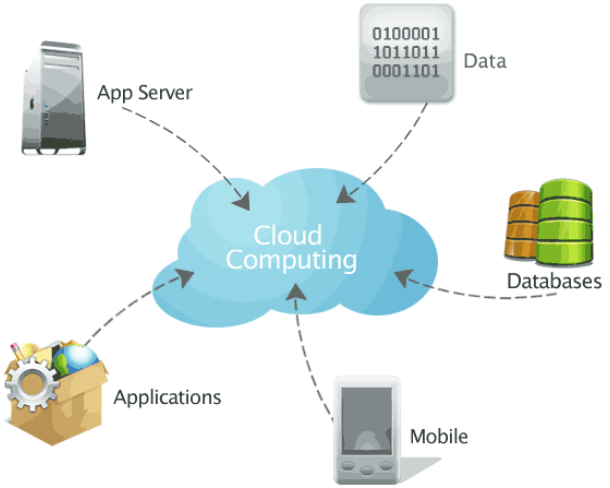
- data security concern

- selecting the perfect cloud set-up

- dependency on service providers

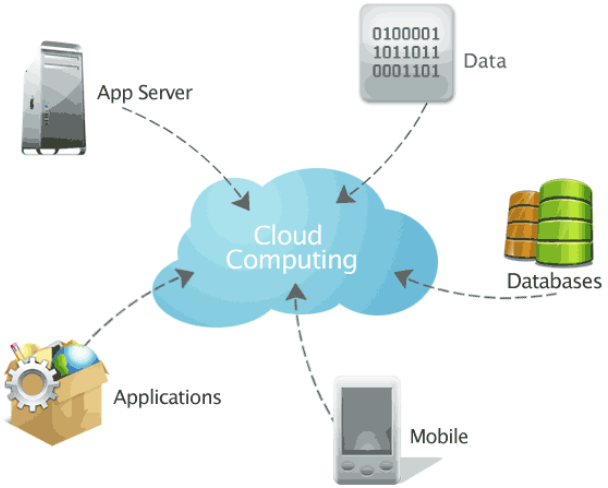
- lack of knowledge and expertise

- cost barrier, consumption basis services charge



avoid **cloud** or/and **web based solutions** as long as possible

Google Cloud

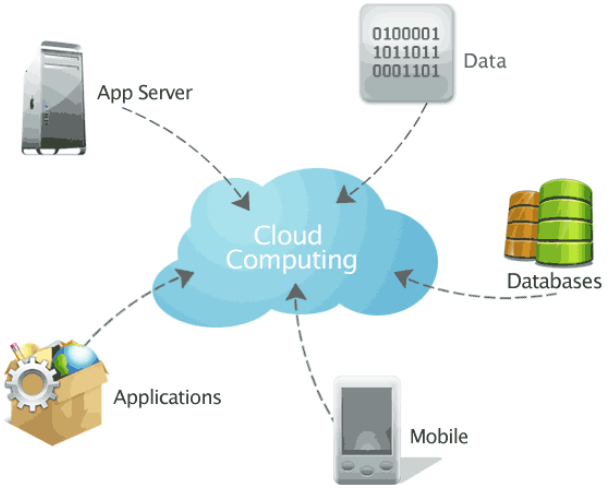


avoid **cloud** or/and **web based solutions** as long as possible

Google Cloud

gcloud CLI

Google Cloud CLI



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

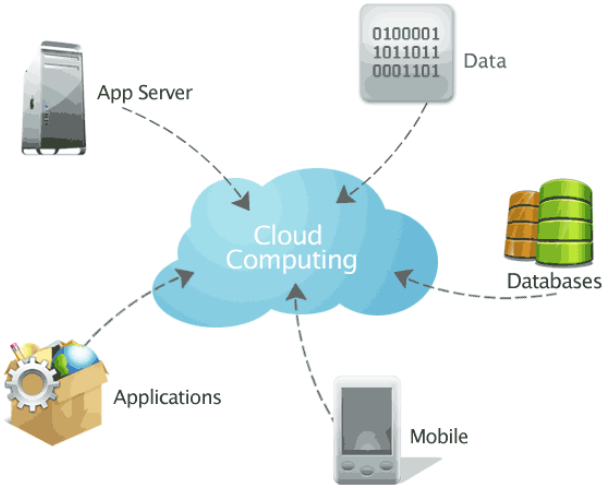
- selecting the perfect cloud set-up

- dependency on service providers

- lack of knowledge and expertise

- cost barrier, consumption basis services charge

- recovery of lost data



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

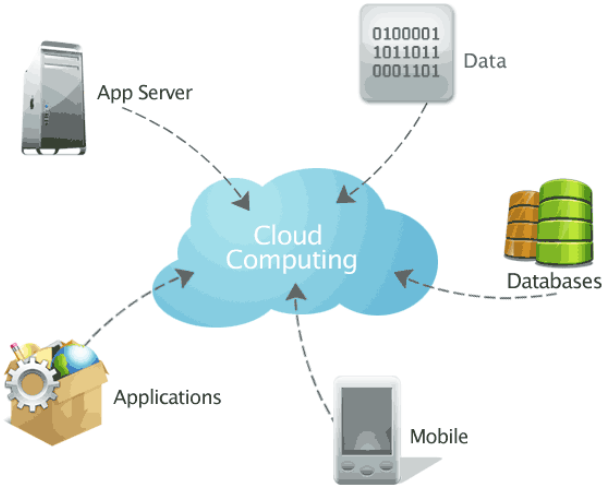
- dependency on service providers

- lack of knowledge and expertise

- cost barrier, consumption basis services charge

- recovery of lost data

- data portability



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

- dependency on service providers

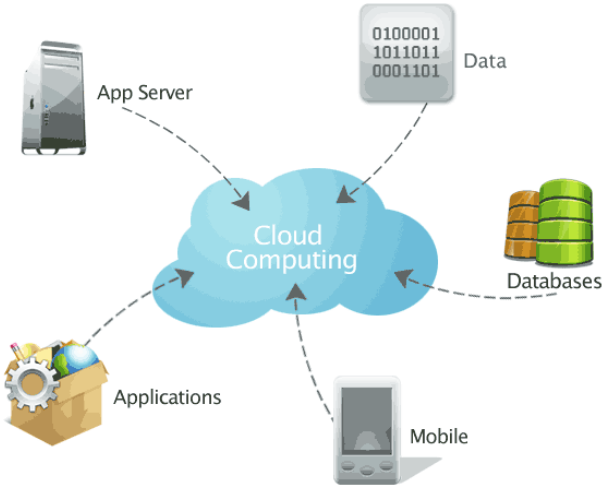
- lack of knowledge and expertise

- cost barrier, consumption basis services charge

- recovery of lost data

- data portability

- hacking



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

- dependency on service providers

- lack of knowledge and expertise

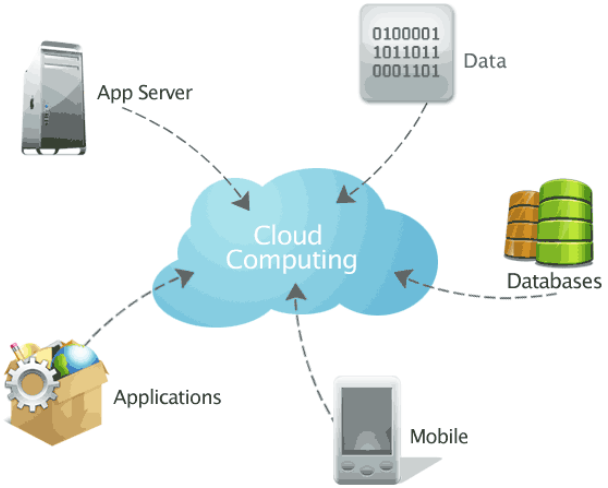
- cost barrier, consumption basis services charge

- recovery of lost data

- data portability

- hacking

- cloud management



avoid **cloud** or/and **web based solutions** as long as possible

You are loosing the control
(no root, no control on data and resources)

- data security concern

- selecting the perfect cloud set-up

- dependency on service providers

- lack of knowledge and expertise

- cost barrier, consumption basis services charge

- recovery of lost data

- data portability

- hacking

- cloud management

- transparency of service provider

avoid **cloud** or/and **web bas**

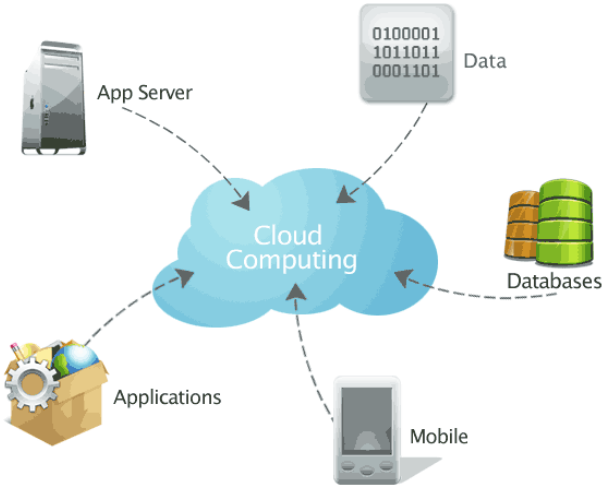
FORBES > INNOVATION

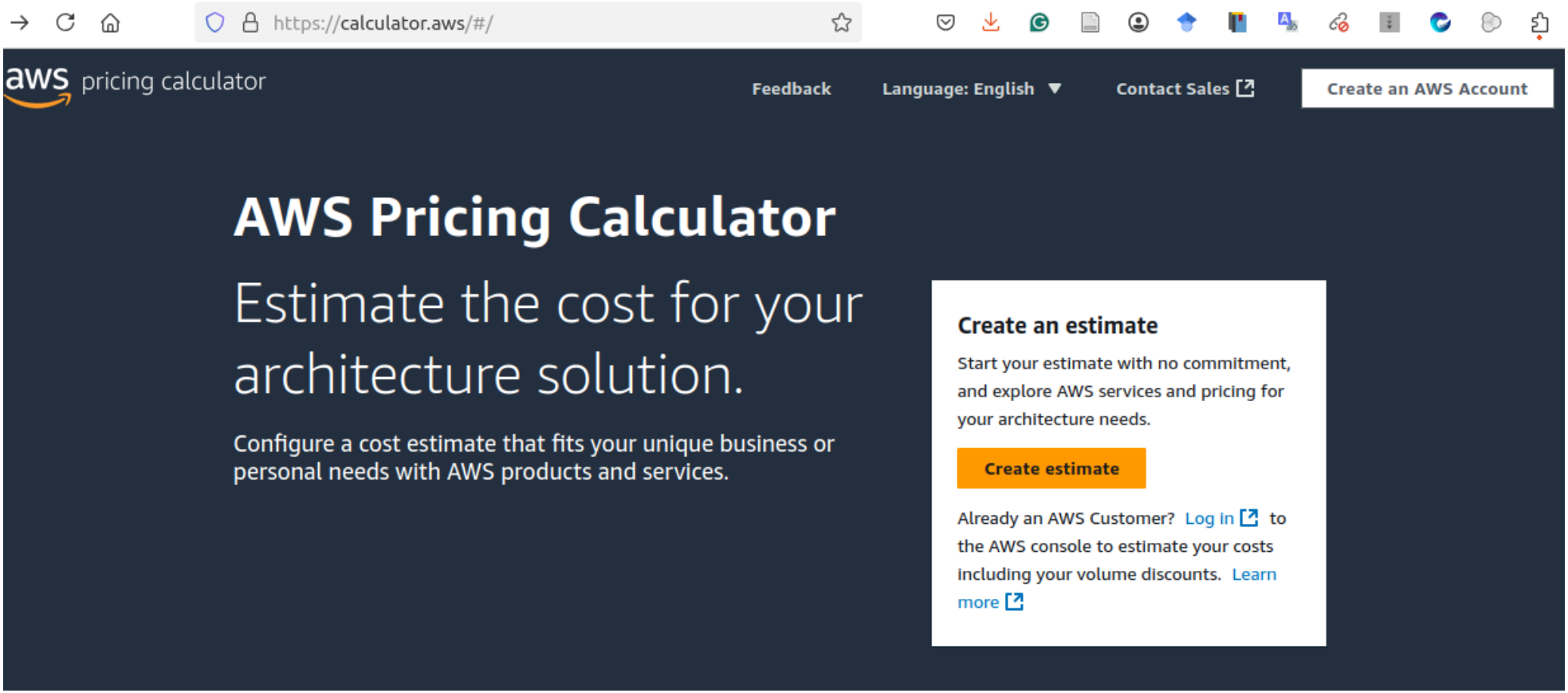
Jeff Bezos Is No Longer The Richest Person In The World After Forgetting To Shut Off EC2 Instance

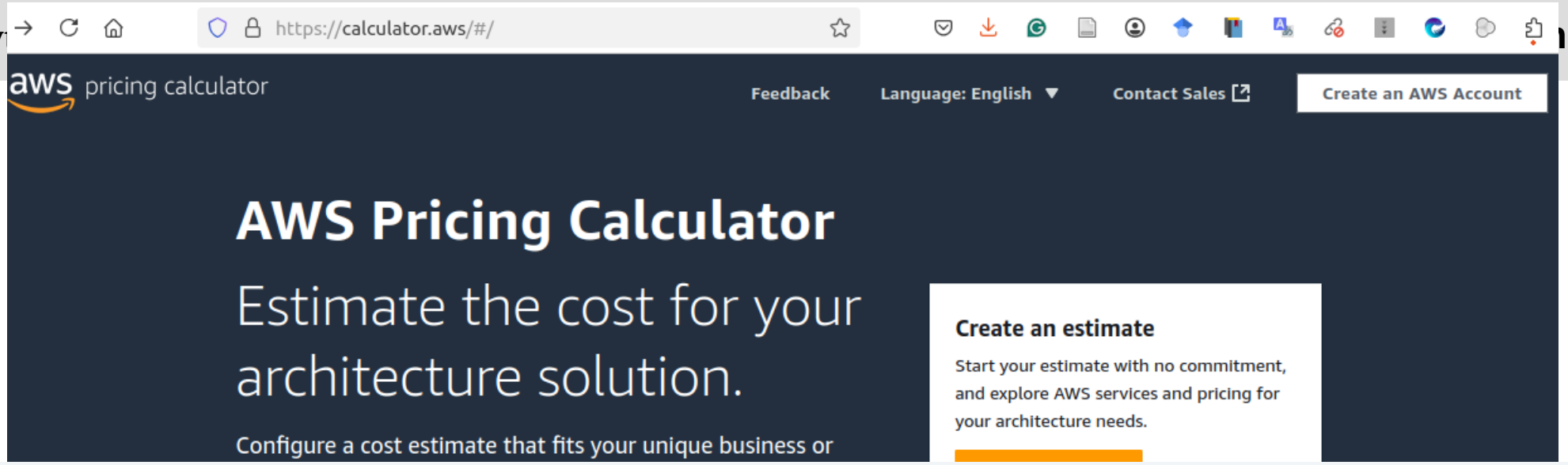
Angel Au-Yeung Former Staff

Updated Oct 25, 2019, 04:54pm EDT

Amazon founder Jeff Bezos (Photo by Jim WATSON / AFP) AFP/GETTY IMAGES

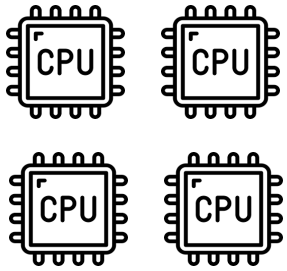

0100001
1011011
0001101
Data
App Server
Cloud
Computing
Databases
Applications
Mobile


https://calculator.aws/#/
aws pricing calculator
Feedback
Language: English
Contact Sales
Create an AWS Account
AWS Pricing Calculator
Estimate the cost for your architecture solution.
Configure a cost estimate that fits your unique business or personal needs with AWS products and services.
Create an estimate
Start your estimate with no commitment, and explore AWS services and pricing for your architecture needs.
Create estimate
Already an AWS Customer? Log in to the AWS console to estimate your costs including your volume discounts. Learn more



	Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)	On Demand Price/hr*	1-yr ISP Effective Hourly (Linux)	3-yr ISP Effective Hourly (Linux)
Single GPU VMs	g6.xlarge	1	24	4	16	1x250	Up to 10	Up to 5	$0.805	$0.499	$0.342
	g6.2xlarge	1	24	8	32	1x450	Up to 10	Up to 5	$0.978	$0.606	$0.416
	g6.4xlarge	1	24	16	64	1x600	Up to 25	8	$1.323	$0.820	$0.562
	g6.8xlarge	1	24	32	128	2x450	25	16	$2.014	$1.249	$0.856
	g6.16xlarge	1	24	64	256	2x940	25	20	$3.397	$2.106	$1.443
	gr6.4xlarge	1	24	16	128	1x600	Up to 25	8	$1.539	$0.954	$0.654
	gr6.8xlarge	1	24	32	256	2x450	25	16	$2.446	$1.517	$1.040
Multi GPU VMs	g6.12xlarge	4	96	48	192	4x940	40	20	$4.602	$2.853	$1.955
	g6.24xlarge	4	96	96	384	4x940	50	30	$6.675	$4.139	$2.837
	g6.48xlarge	8	192	192	768	8x940	100	60	$13.35	$8.277	$5.674

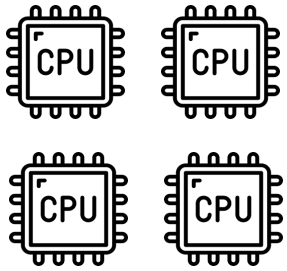
HPC vs Cloud



~0.096 CPUh

0.70-0.96 GPUh

HPC vs Cloud



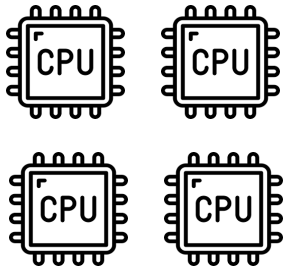
~0.096 CPUh

0.70-0.96 GPUh

In my last bioinformatics project I used:

- **1M GPUh**
- **15M CPUh**

HPC vs Cloud



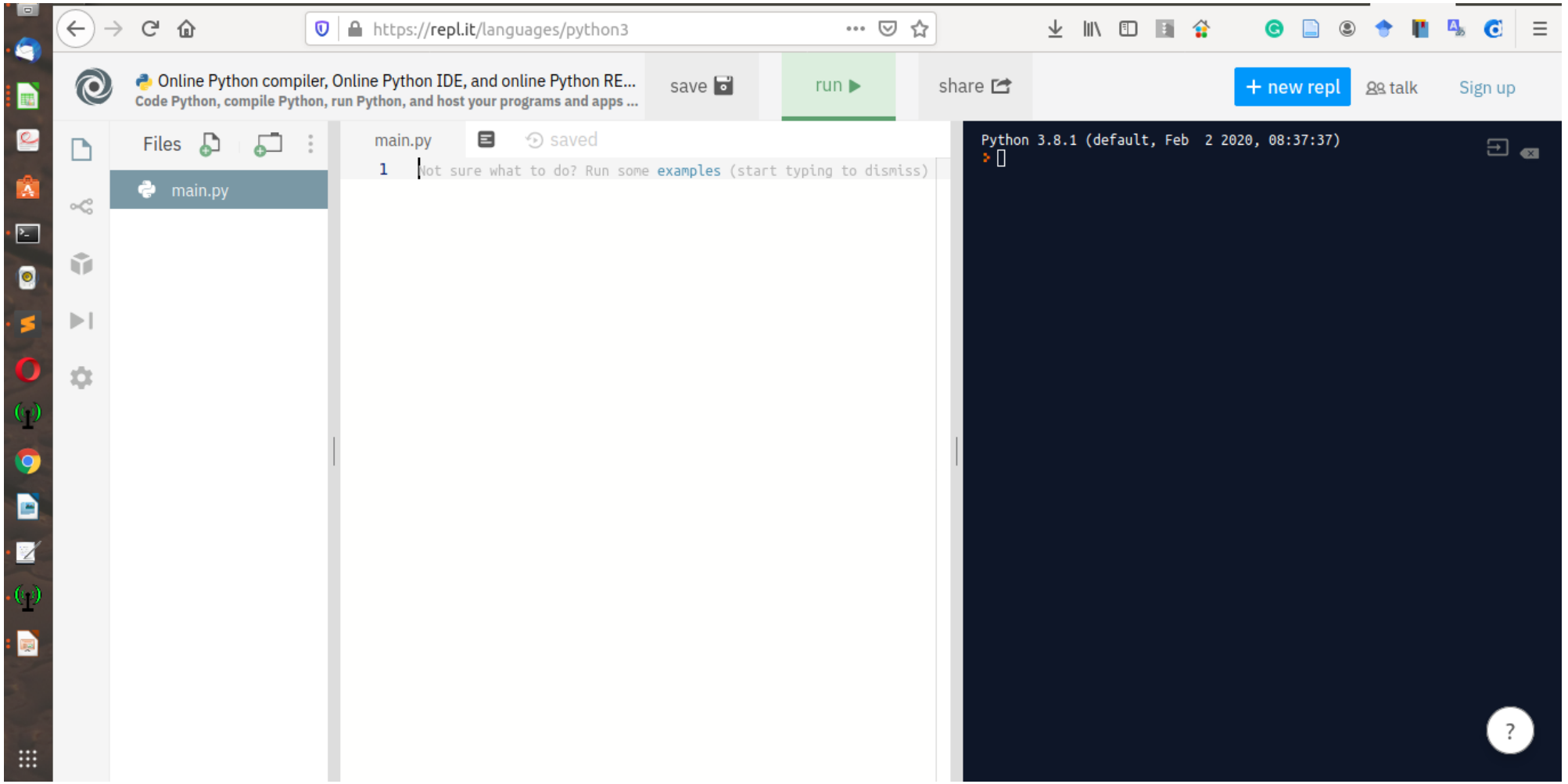
~0.096 CPUh

0.70-0.96 GPUh

In my last bioinformatics project I used:

- **1M GPUh**
- **15M CPUh**

1M * 0.7 + 15M * 0.096 = $2.14M



NEVER EVER USE
THE WEB BROWSER ONLINE PYTHON EDITORS
DURING LABS

Data Visualization starts with

Data Visualization starts with the DATA

Data Visualization starts with the DATA

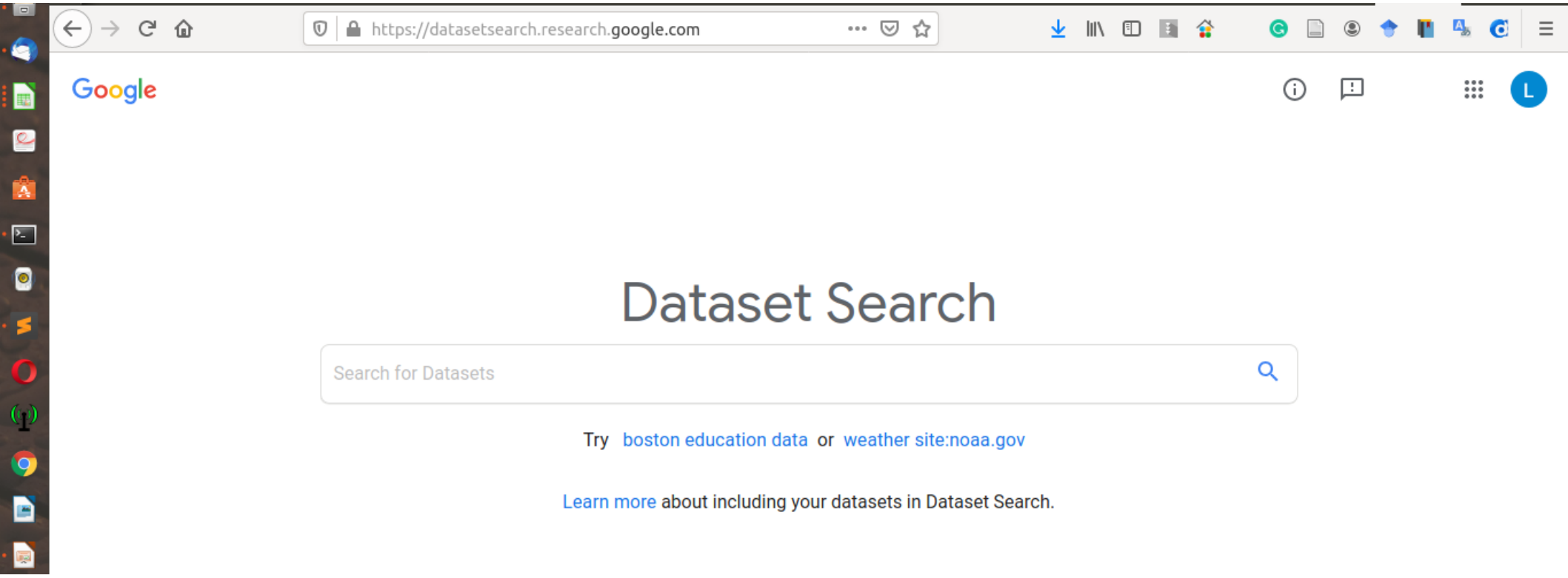
- Data Acquisition
- Data Aggregation
- Data Verification
- Data Validation
- Data Redundancy
- Data Mining
- Data Recovery
- Data Integrity

Data Visualization starts with the DATA

- Data Acquisition
- Data Aggregation
- Data Verification
- Data Validation
- Data Redundancy
- Data Mining
- Data Recovery
- Data Integrity

- Acquire
- Parse
- Filter
- Mine
- Represent
- Refine
- Interact

Data Visualization starts with the DATA



https://datasetsearch.research.google.com/

Data Visualization starts with the DATA

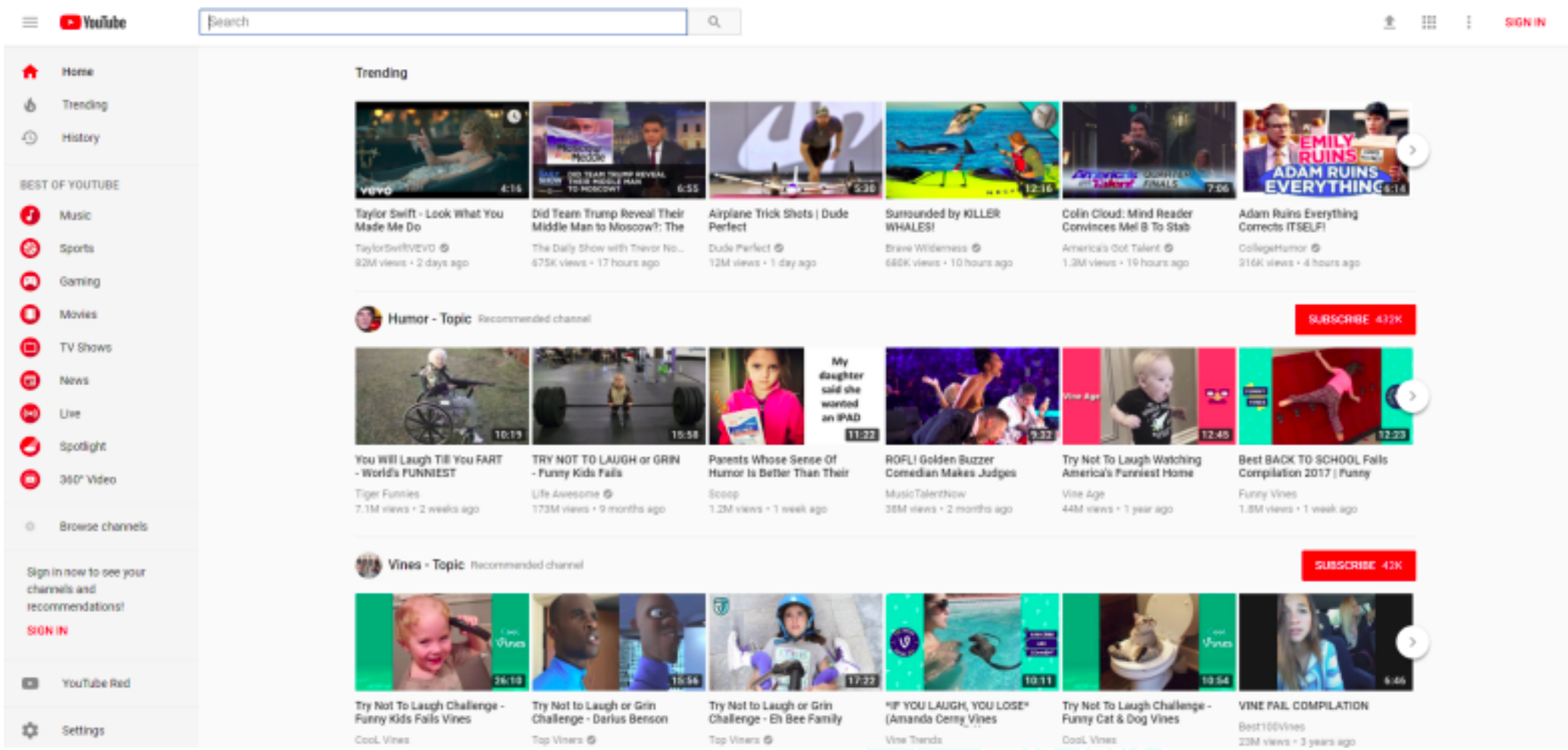
Data Visualization starts with the DATA

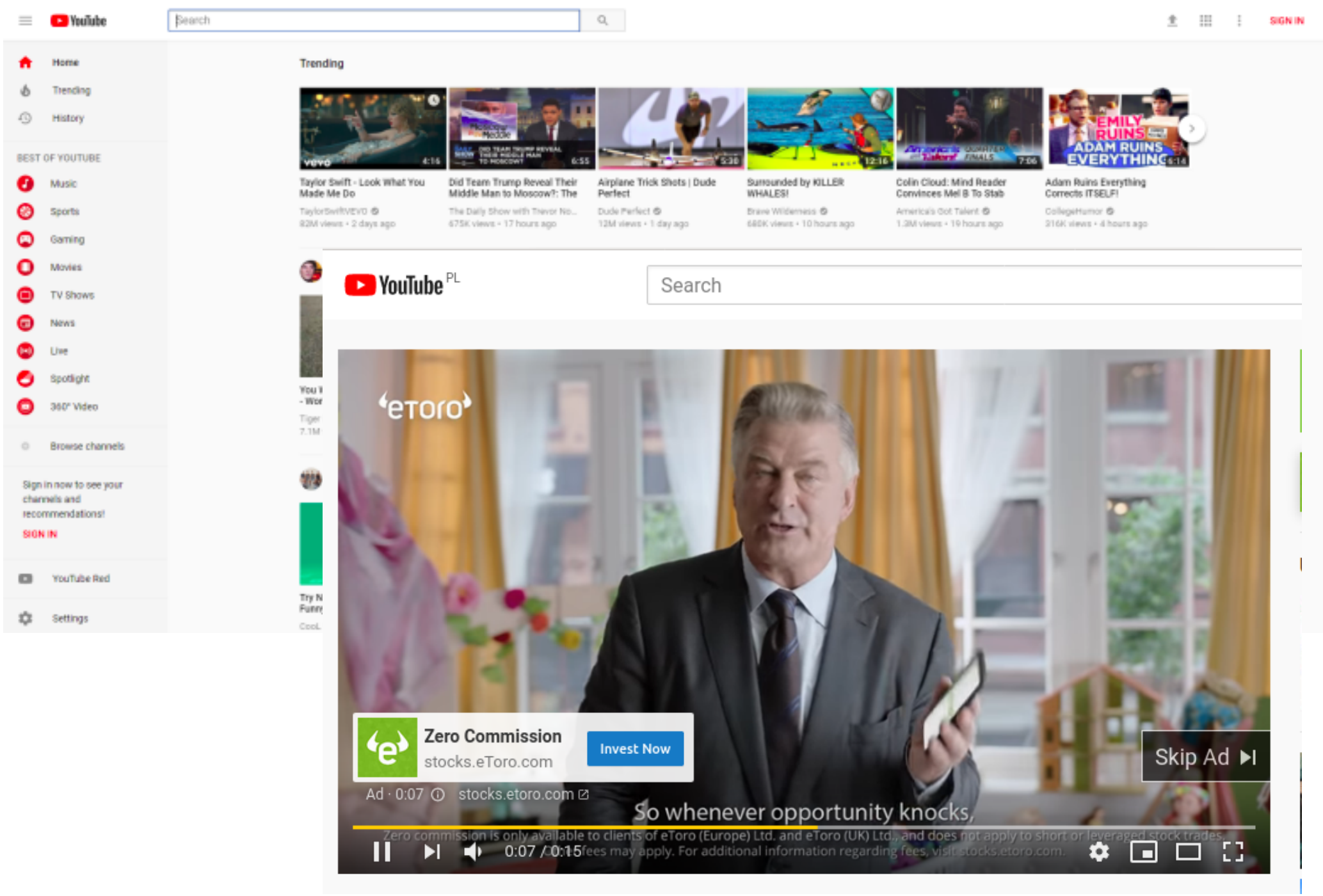
Data Visualization starts with the DATA

Data Visualization starts with the DATA

Google bought **YouTube** in 2006

for US$1.65 billion




YouTube
Search
SIGN IN
Home
Trending
History
BEST OF YOUTUBE
Music
Sports
Gaming
Movies
TV Shows
News
Live
Spotlight
360° Video
Browse channels
Sign in now to see your channels and recommendations!
SIGN IN
YouTube Red
Settings
Trending
Taylor Swift - Look What You Made Me Do
Did Team Trump Reveal Their Middle Man to Moscow?: The
Airplane Trick Shots | Dude Perfect
Surrounded by KILLER WHALES!
Colin Cloud: Mind Reader Convinces Mel B To Stab
Adam Ruins Everything Corrects ITSELF!
YouTube PL
Search
eToro
Zero Commission
stocks.eToro.com
Invest Now
Ad · 0:07
stocks.etoro.com
Skip Ad
So whenever opportunity knocks,
Zero commission is only available to clients of eToro (Europe) Ltd. and eToro (UK) Ltd., and does not apply to short or leveraged stock trades.
fees may apply. For additional information regarding fees, visit stocks.etoro.com.
0:07 / 0:15

Data Visualization starts with the DATA

Data Visualization starts with the DATA

Data Visualization starts with the DATA

Cambridge Analytica had improperly obtained the private information of 87 million Facebook users.

Data Visualization starts with the DATA

The personal data of up to **87 million** Facebook users were acquired **via** the **270,000 Facebook users** who used a Facebook **app** called "**This Is Your Digital Life**." By giving this third-party app permission to acquire their data, back in 2015, this also gave the app access to information on the user's friends network; this resulted in the data of about 87 million users, the majority of whom had not explicitly given Cambridge Analytica permission to access their data, being collected. The app developer breached Facebook's terms of service by giving the data to Cambridge Analytica.

Wikipedia

Data Visualization starts with the DATA

In 2016 used for Donald Trump's presidential campaign

Data Visualization starts with the DATA

In 2016 used for Donald Trump's presidential campaign

Leave.EU (one of the organisations campaigning in the UKs referendum on Brexit)

Data Visualization starts with the DATA

In 2016 used for Donald Trump's presidential campaign

Leave.EU (one of the organisations campaigning in the UKs referendum on Brexit)

The company was involved in 44 US political races in 2014

Data Visualization starts with the DATA

Microsoft Azure

Closed May 1 2018

ORACLE

INTERNET NEWS DECEMBER 6, 2019 / 6:34 PM / 3 MONTHS AGO

Hungary watchdog fines Facebook for misleading users

1 MIN READ

INTERNET NEWS DECEMBER 6, 2019 / 6:34 PM / 3 MONTHS AGO

Hungary watchdog fines Facebook for misleading users

BUDAPEST (Reuters) - Hungary’s competition watchdog has fined Facebook 1.2 billion forints ($4 million), its biggest fine to date, for misleading users by claiming its services were free.

General Data Protection Regulation
a.k.a RODO

General Data Protection Regulation
a.k.a RODO

Public Domain

Creative Commons (CC)

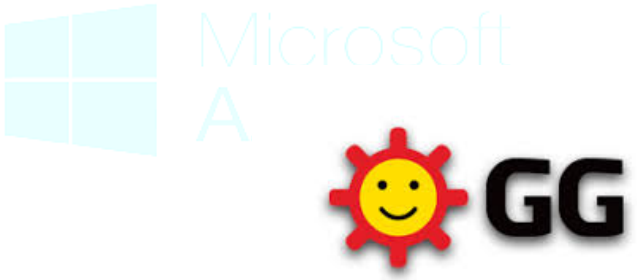
MIT License

GNU General Public License (GPL)

Property licenses

Data Visualization starts with the DATA

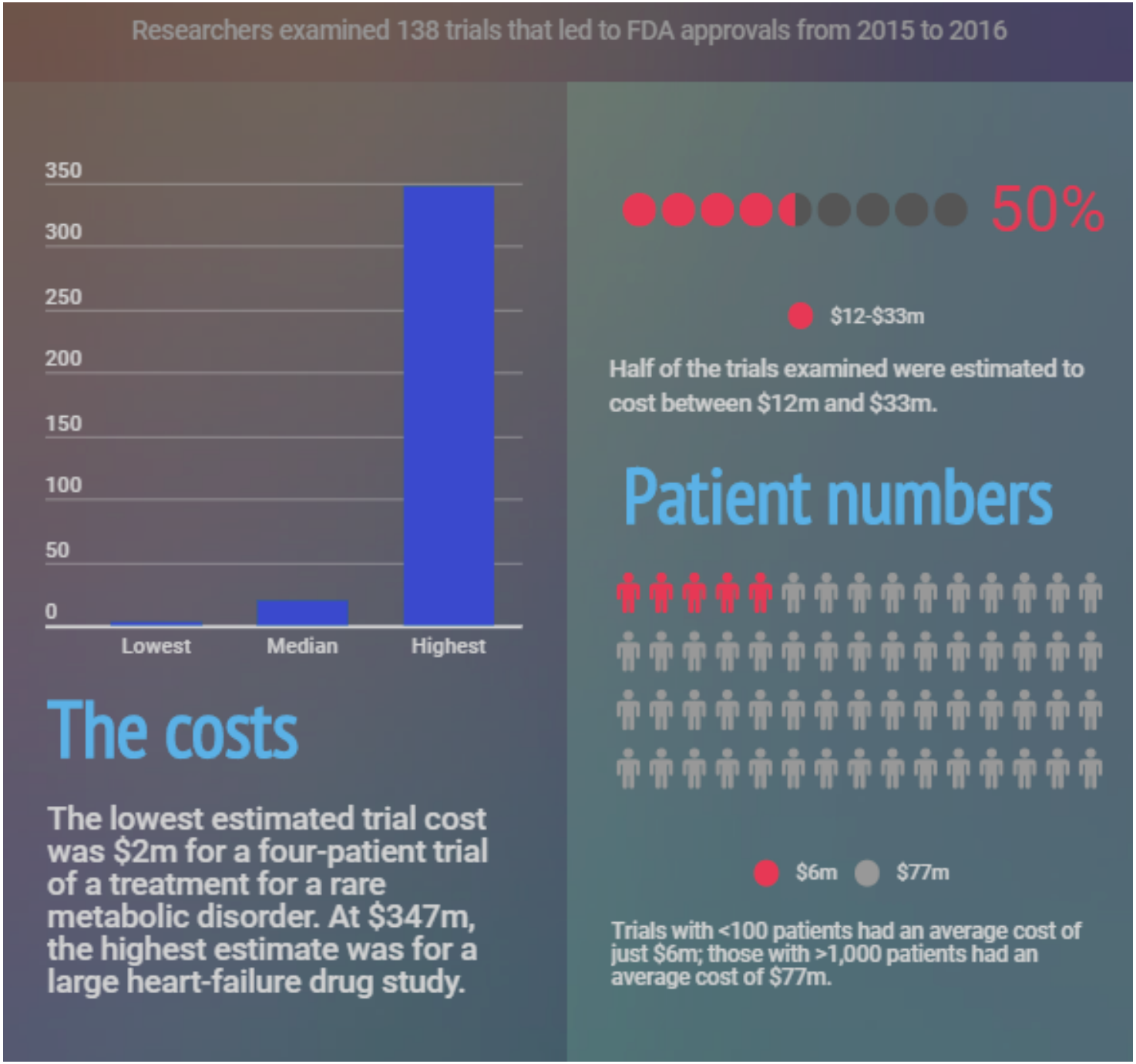
Data Visualization starts with the DATA



General information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people.

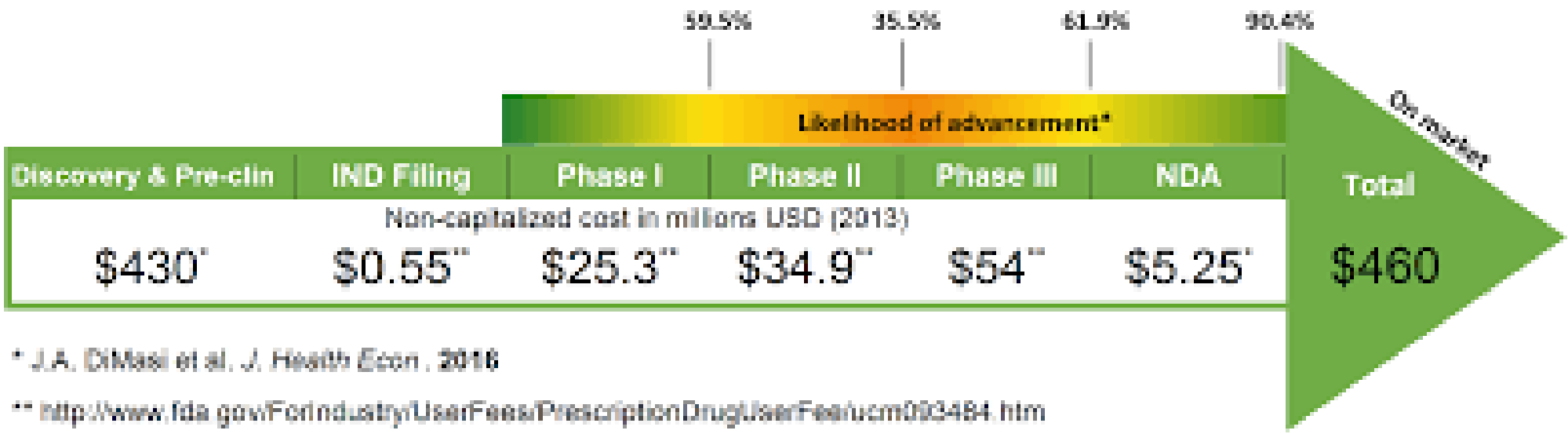
General information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people.

… but if the data are in the context they can cost much more (and the sky is the limit)


Researchers examined 138 trials that led to FDA approvals from 2015 to 2016
350
300
250
200
150
100
50
0
Lowest
Median
Highest
The costs
The lowest estimated trial cost was $2m for a four-patient trial of a treatment for a rare metabolic disorder. At $347m, the highest estimate was for a large heart-failure drug study.
50%
$12-$33m
Half of the trials examined were estimated to cost between $12m and $33m.
Patient numbers
$6m
$77m
Trials with <100 patients had an average cost of just $6m; those with >1,000 patients had an average cost of $77m.

General information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people.

… but if the data are <u>in the context</u> they can cost much more (and the sky is the limit)



kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

Kaggle is the place to do data science projects

See how it works

Sign up with just one click:

We won't share anything without your permission

Google Facebook Yahoo

Manually create an account:

Email

Password

Sign Up

Start a new project

New Project

Explore projects created by others

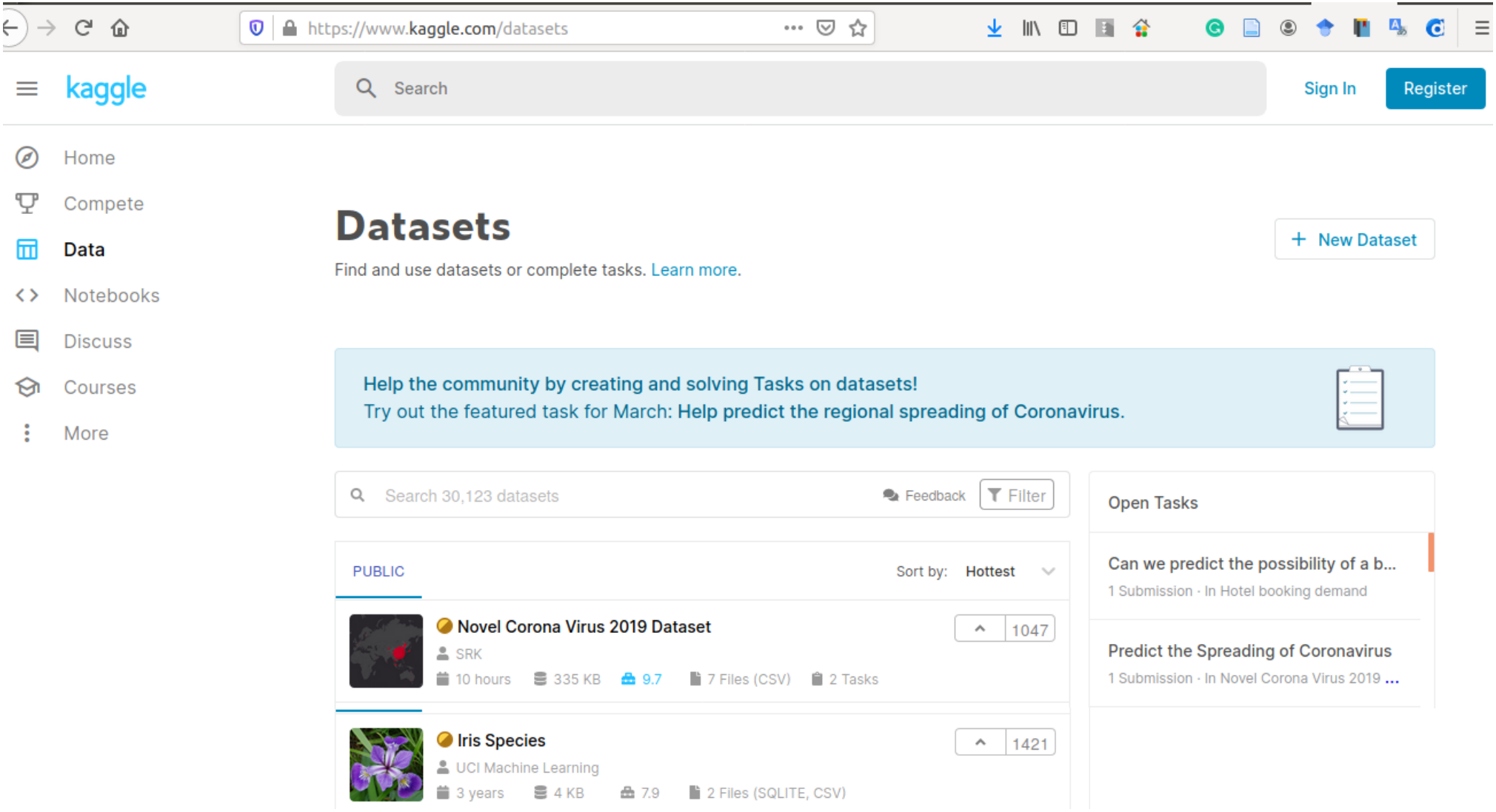
View Projects

Join one of our competitions

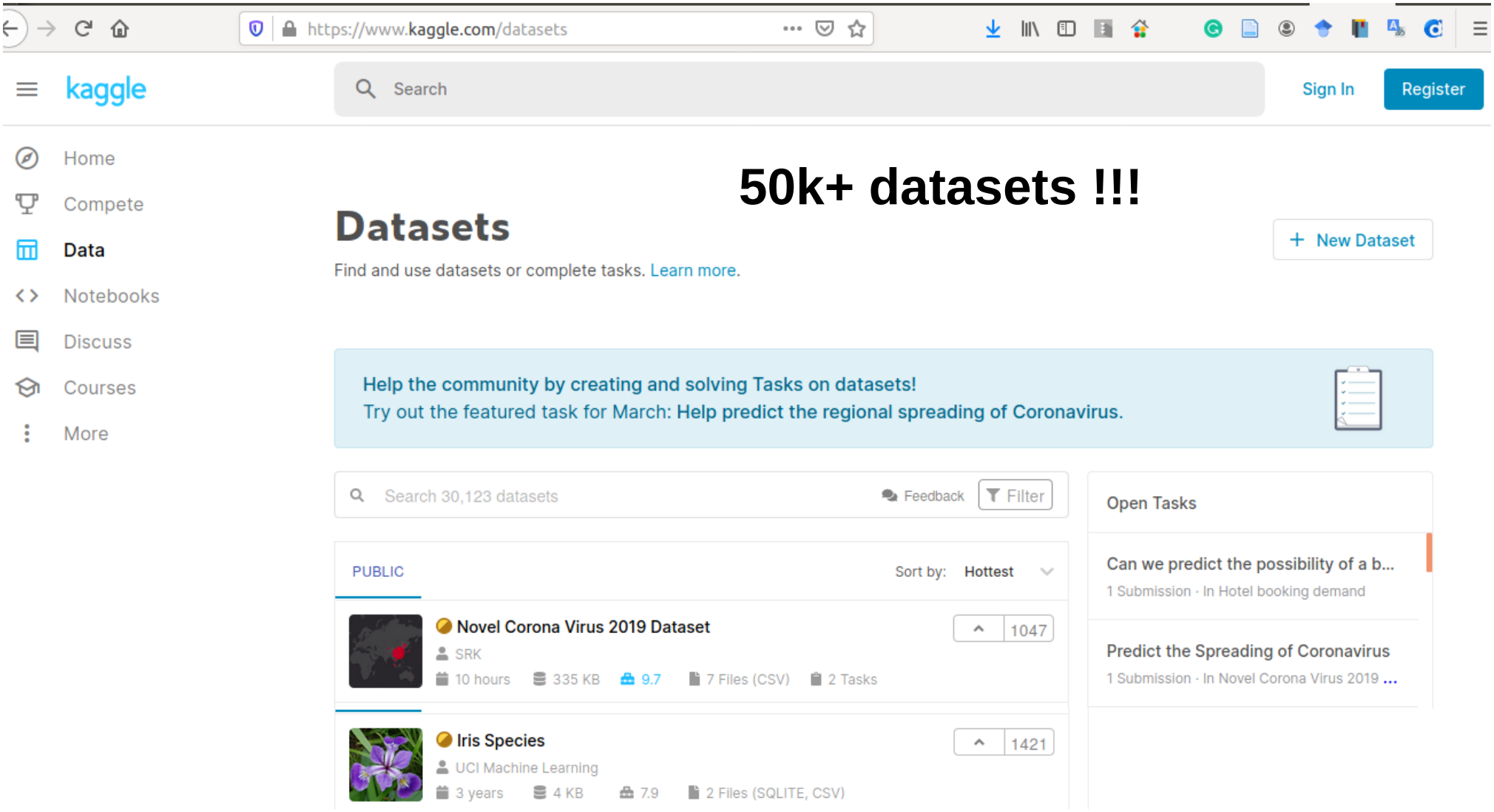
Join Competitions

h

kaggle


https://www.kaggle.com/datasets
kaggle
Search
Sign In
Register
Home
Compete
Data
Notebooks
Discuss
Courses
More
Datasets
Find and use datasets or complete tasks. Learn more.
+ New Dataset
Help the community by creating and solving Tasks on datasets!
Try out the featured task for March: Help predict the regional spreading of Coronavirus.
Search 30,123 datasets
Feedback
Filter
PUBLIC
Sort by: Hottest
Novel Corona Virus 2019 Dataset
SRK
10 hours
335 KB
9.7
7 Files (CSV)
2 Tasks
1047
Iris Species
UCI Machine Learning
3 years
4 KB
7.9
2 Files (SQLITE, CSV)
1421
Open Tasks
Can we predict the possibility of a b...
1 Submission · In Hotel booking demand
Predict the Spreading of Coronavirus
1 Submission · In Novel Corona Virus 2019 ...

kaggle


https://www.kaggle.com/datasets
kaggle
Search
Sign In
Register
Home
Compete
Data
Notebooks
Discuss
Courses
More
50k+ datasets !!!
Datasets
Find and use datasets or complete tasks. Learn more.
+ New Dataset
Help the community by creating and solving Tasks on datasets!
Try out the featured task for March: Help predict the regional spreading of Coronavirus.
Search 30,123 datasets
Feedback
Filter
PUBLIC
Sort by: Hottest
Novel Corona Virus 2019 Dataset
SRK
10 hours
335 KB
9.7
7 Files (CSV)
2 Tasks
1047
Iris Species
UCI Machine Learning
3 years
4 KB
7.9
2 Files (SQLITE, CSV)
1421
Open Tasks
Can we predict the possibility of a b...
1 Submission · In Hotel booking demand
Predict the Spreading of Coronavirus
1 Submission · In Novel Corona Virus 2019 ...


Główny
Urząd Statystyczny


U.S. BUREAU OF LABOR STATISTICS


United States®
Census
Bureau


THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP


World Health
Organization


eurostat


UNESCO

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

l.kozlowski@mimuw.edu.pl