UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

IMD0905 - Data Science I

Lesson #2 - Data Science Platforms

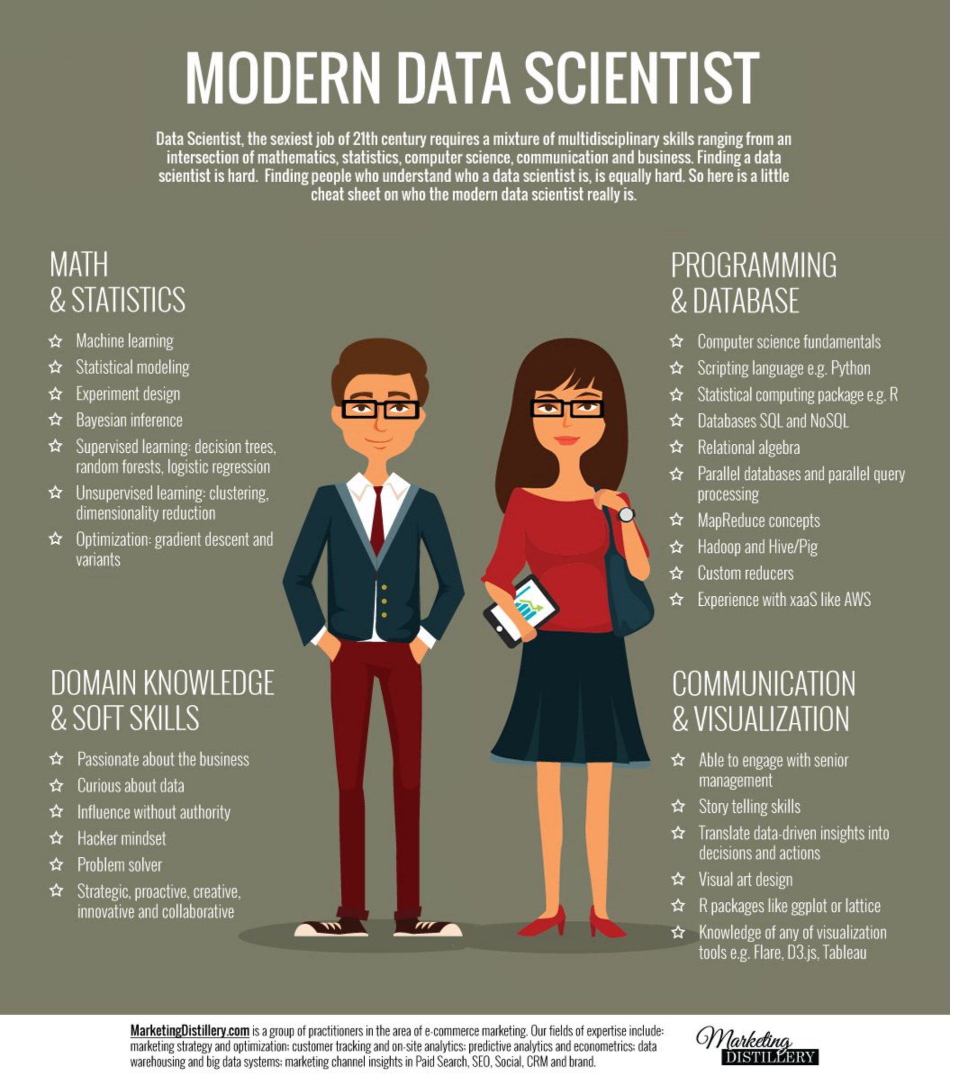
Ivanovitch Silva

August, 2018

Agenda

- How to Become a Data Scientist
- Development platform
- Hello World
- Python Beginner

How to Become a **Data Scientist**



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

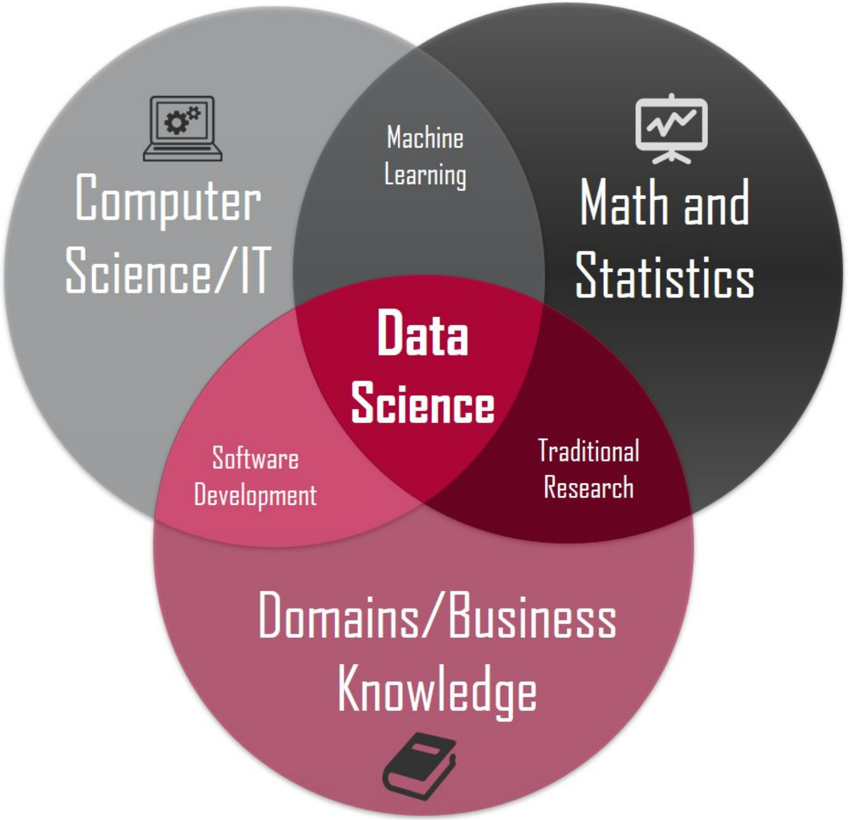
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

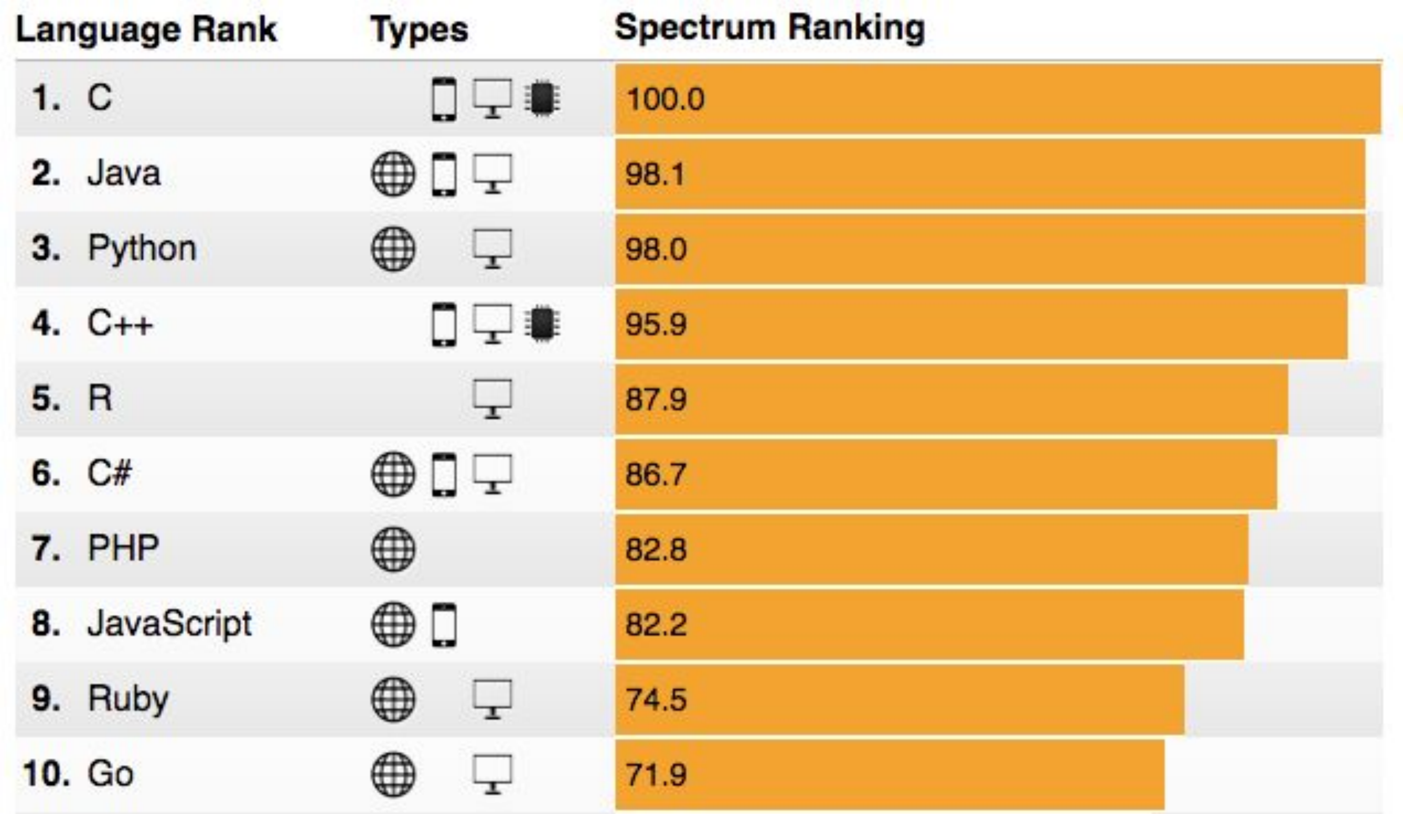
Marketing DISTILLERY



Pick ONE programming language and STICK to it. Don’t go back and constantly change your choice of language to study. If you do, you will slow your progress down.

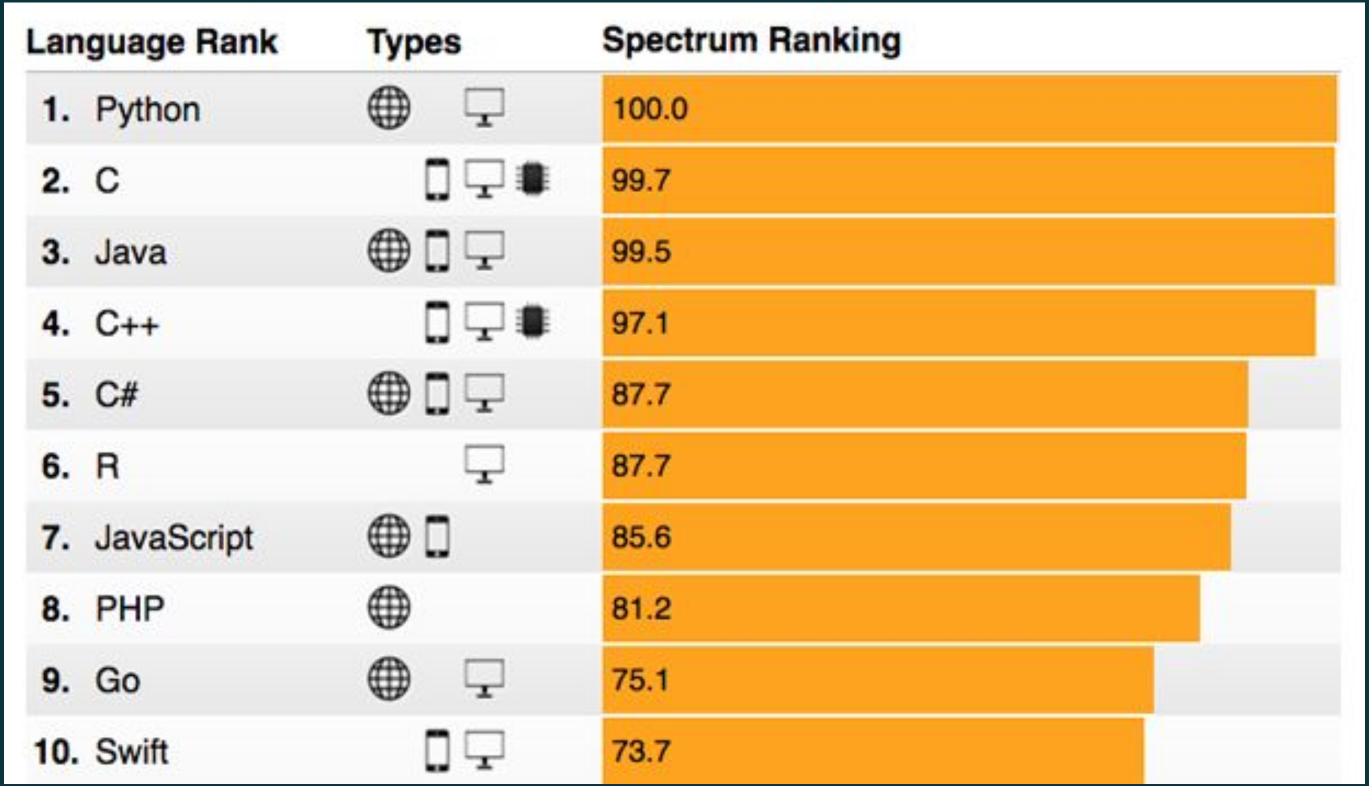
which programming language to learn first (DS)?

https://goo.gl/VKYfXn



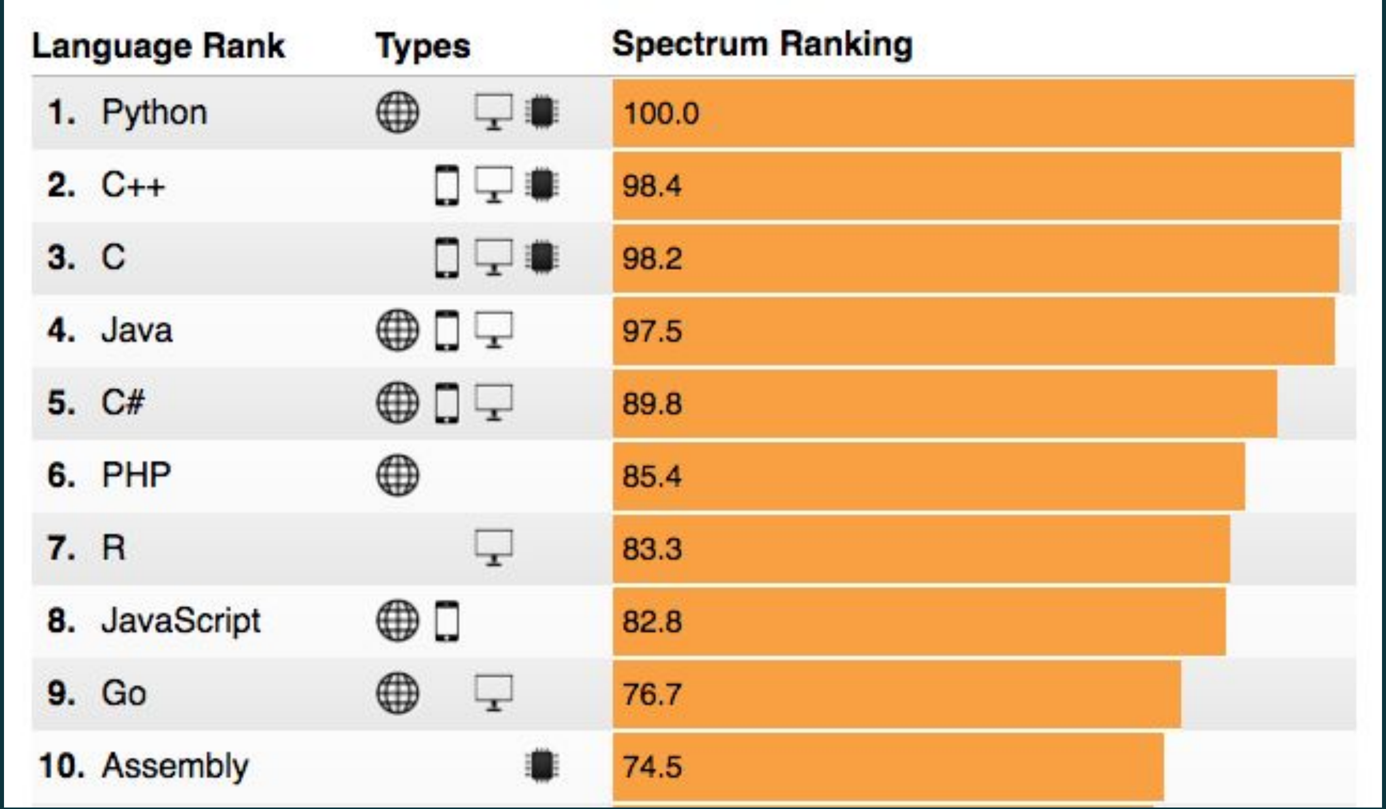
Language Rank	Types	Spectrum Ranking
1. C	📱 🖥 ▣	100.0
2. Java	🌐 📱 🖥	98.1
3. Python	🌐 🖥	98.0
4. C++	📱 🖥 ▣	95.9
5. R	🖥	87.9
6. C#	🌐 📱 🖥	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐 📱	82.2
9. Ruby	🌐 🖥	74.5
10. Go	🌐 🖥	71.9

IEEE Spectrum - Jul 2016 https://goo.gl/BqrkDI



Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥	100.0
2. C	📱 🖥 ▦	99.7
3. Java	🌐 📱 🖥	99.5
4. C++	📱 🖥 ▦	97.1
5. C#	🌐 📱 🖥	87.7
6. R	🖥	87.7
7. JavaScript	🌐 📱	85.6
8. PHP	🌐	81.2
9. Go	🌐 🖥	75.1
10. Swift	📱 🖥	73.7

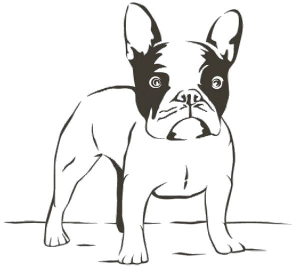
IEEE Spectrum - Jul 2017 https://goo.gl/HSPLWe



Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥 ■	100.0
2. C++	📱 🖥 ■	98.4
3. C	📱 🖥 ■	98.2
4. Java	🌐 📱 🖥	97.5
5. C#	🌐 📱 🖥	89.8
6. PHP	🌐	85.4
7. R	🖥	83.3
8. JavaScript	🌐 📱	82.8
9. Go	🌐 🖥	76.7
10. Assembly	■	74.5

IEEE Spectrum - Jul 2018
https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages

have
a pet
project



Be clear about your motivation. The reason this is important because learning Data Science is HARD. VERY HARD! So it’s easy to lose motivation when on the journey.

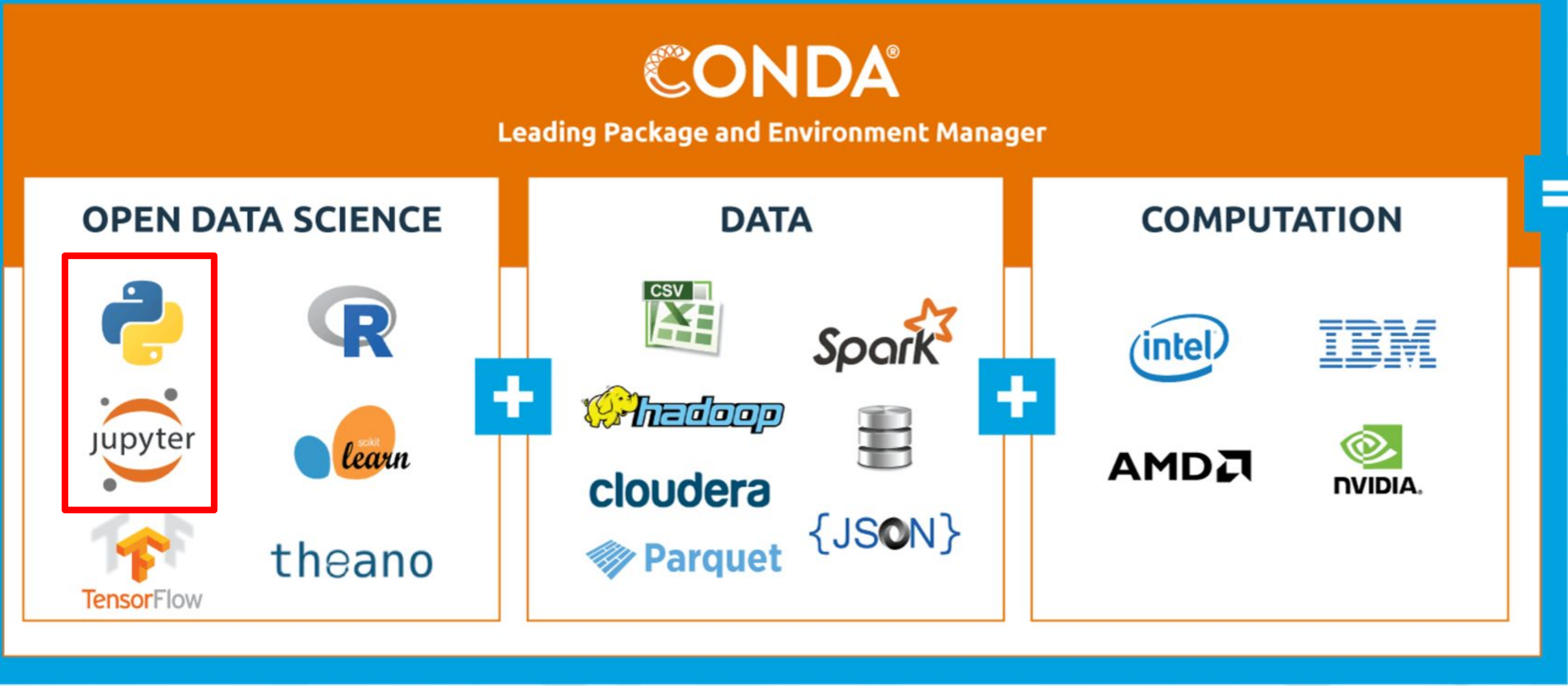
Immerse yourself in the community (newsletters, articles, books, podcasts, youtube, hackathons and meetups)

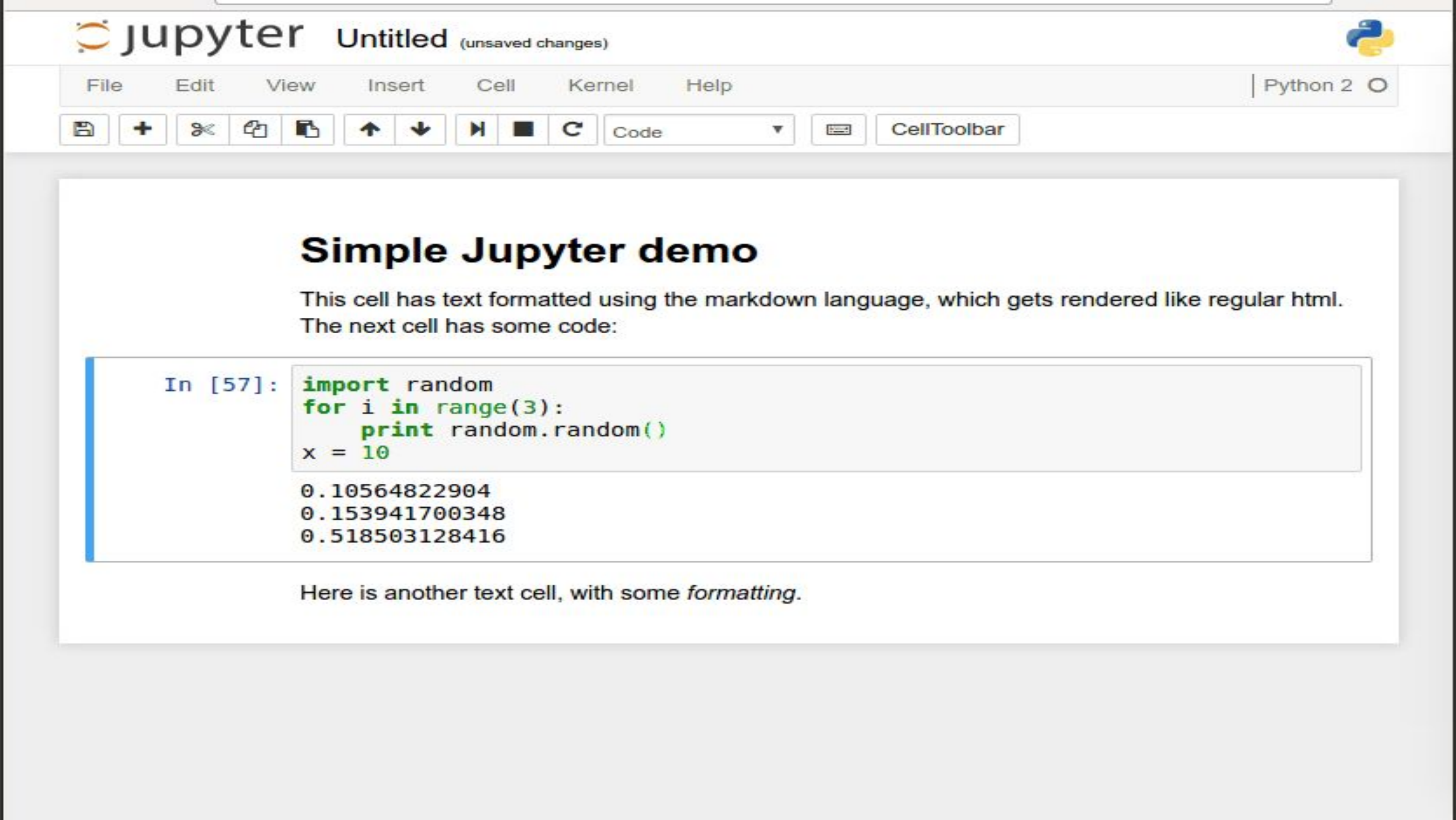
https://www.continuum.io/downloads

Why Anaconda?

ANACONDA

Leading Open Data Science Platform Powered by Python





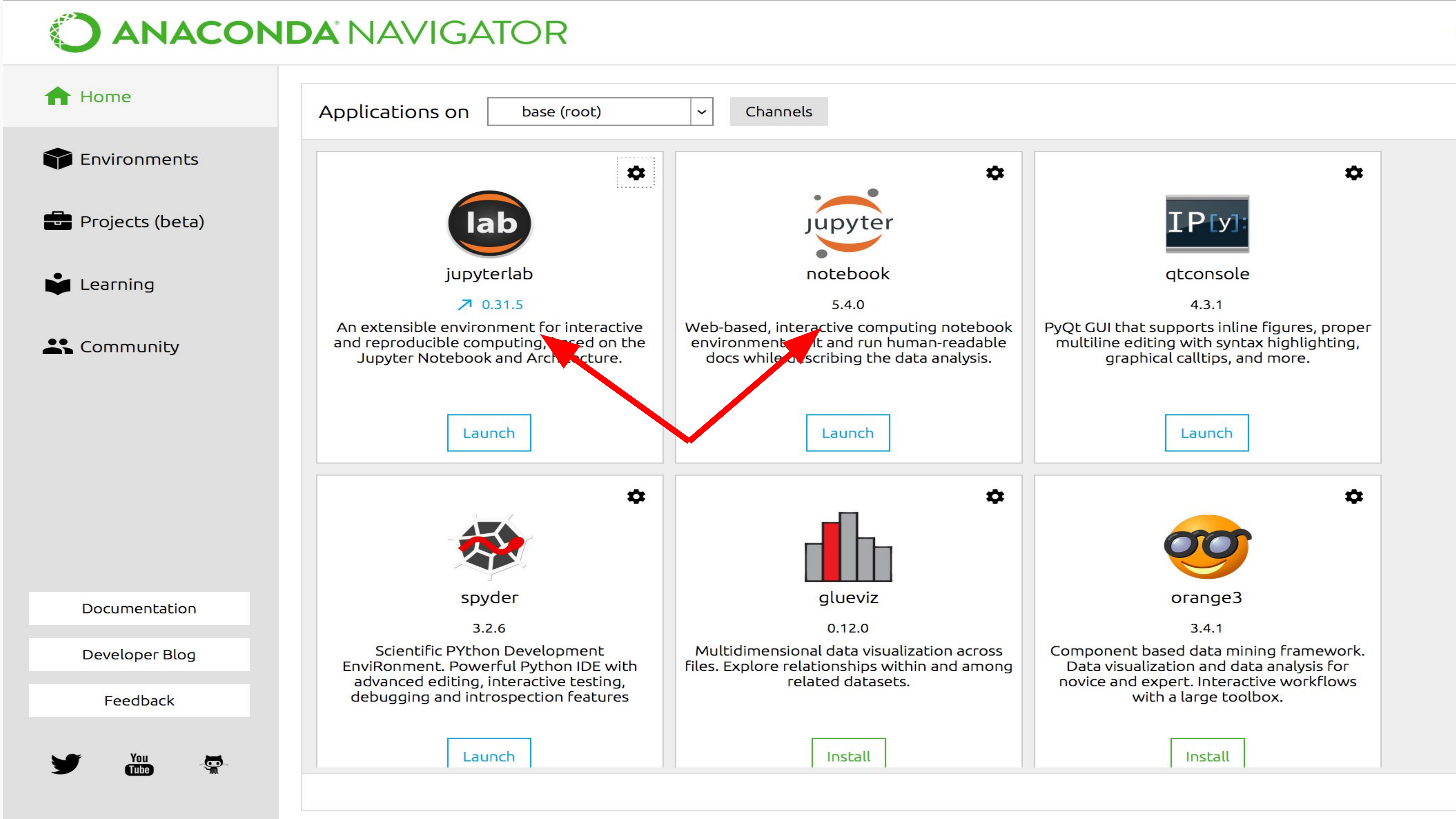
Simple Jupyter demo

This cell has text formatted using the markdown language, which gets rendered like regular html. The next cell has some code:

```
In [57]: import random
         for i in range(3):
             print random.random()
         x = 10
```

```
0.10564822904
0.153941700348
0.518503128416
```

Here is another text cell, with some *formatting*.



An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Sponsors

Project Jupyter receives direct funding from the following sources:

THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST

ALFRED P. SLOAN
FOUNDATION

GORDON AND BETTY
MOORE
FOUNDATION

Google

rackspace®
the #1 managed cloud company

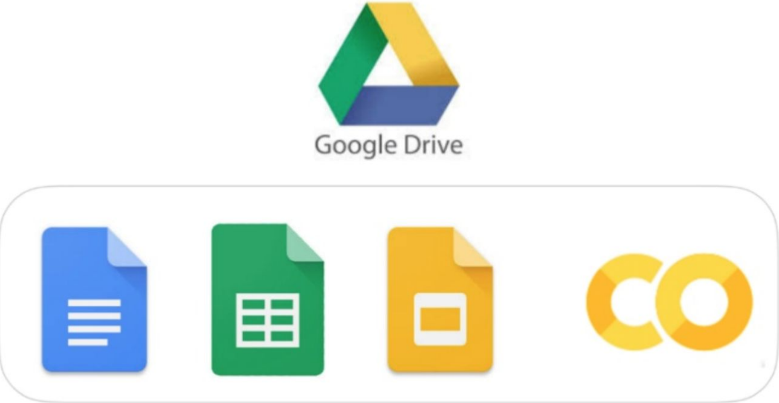
fastly®

Horizon 2020
European Union funding
for Research & Innovation

Microsoft

Google Colaboratory

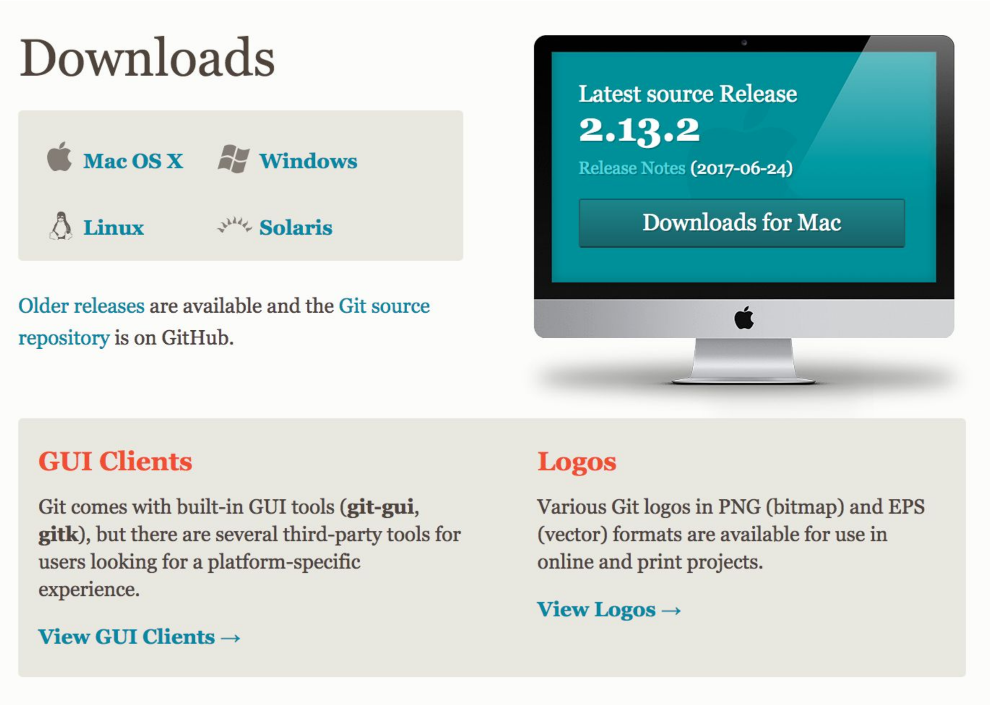
https://colab.research.google.com/



Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.

Colaboratory notebooks are stored in Google Drive and can be shared just as you would with Google Docs or Sheets. Colaboratory is free to use.

Installing Git



Downloads

Latest source Release

2.13.2

Release Notes (2017-06-24)

Older releases are available and the Git source repository is on GitHub.

GUI Clients

Git comes with built-in GUI tools (**git-gui**, **gitk**), but there are several third-party tools for users looking for a platform-specific experience.

View GUI Clients →

Logos

Various Git logos in PNG (bitmap) and EPS (vector) formats are available for use in online and print projects.

View Logos →

https://git-scm.com/downloads

Atualizar o repositório

```
git clone https://github.com/ivanovitchm/IMD0905_datascience_one.git
```

Ou ….

```
git pull
```

```
        console.log("response.data", response.data);
        this.setState({
          ajaxReturn: response.data
        });
      }.bind(this));
    },

    addContactAjaxCall: function() {
      console.log("Add new contact:", this.state.name, this.state.
      const contact = {
        name: this.state.name,
        email: this.state.email
      };
      ajaxHelpers.addContact(contact)
    },

    nameInput: function(e){
      this.setState({
        name: e.target.value
      })
    },
    emailInput: function(e){
      this.setState({
        email: e.target.value
      })
    },
    newNameInput: function(e){
      this.setState({
        newName: e.target.value
      })
    },
    newEmailInput: function(e){
```

```
import React from "react";

const FindContact = React.createClass({

  render: function(){
    return(
      <div className="find-contact">
        <label>Name:</label>
        <input
          type="text"
          className="find-name"
          placeholder="Name"
          onChange={this.props.getSearchName}
          />
        <button
          type="button"
          onClick={this.props.searchAjaxCall}>
          Get Contact
        </button>

      </div>
    )
  }
})

export default FindContact;
```

Python Beginner

- Python basic
- Files and Loops
- Boolean and If statements
- List operations
- Challenges

Notebook: "Lesson #2 - Python Beginner.ipynb"

Lesson #2