

Fist steps with data science

**Thomas Stoeger
Northwestern University
Postdoctoral Association**

What do I need?

What do I need?

What can I ignore?

What do I need?

What can I ignore?

Bonus: install and start an environment

What do I need?

What do I need?

a reason

What do I need?

a reason

tools ready to be used

What do I need?

a reason

tools ready to be used

help, if getting stuck

What do I need?

a reason

tools ready to be used

help, if getting stuck



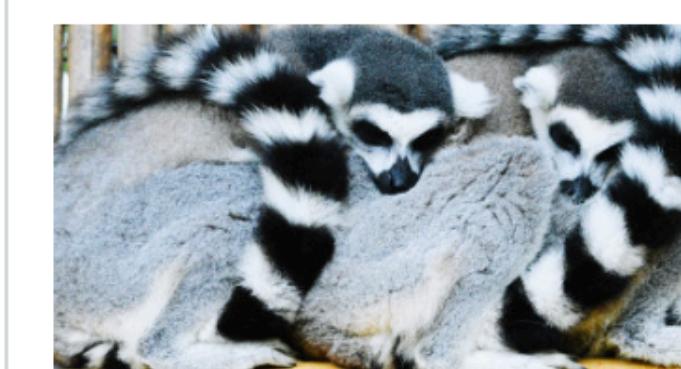
Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen
on a page from a high school yearbook, 8.5" x 12"

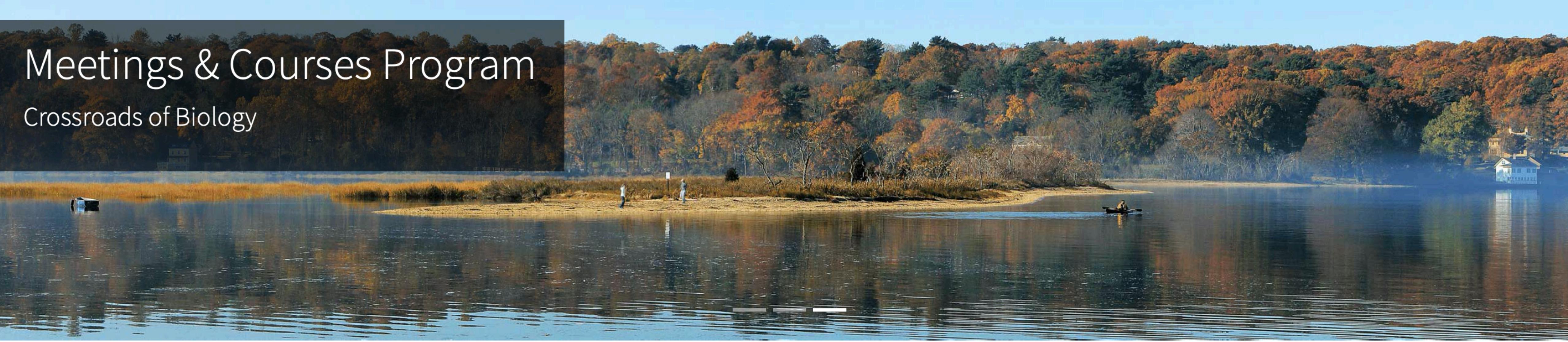
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

WHAT TO READ NEXT





Meetings & Courses Program

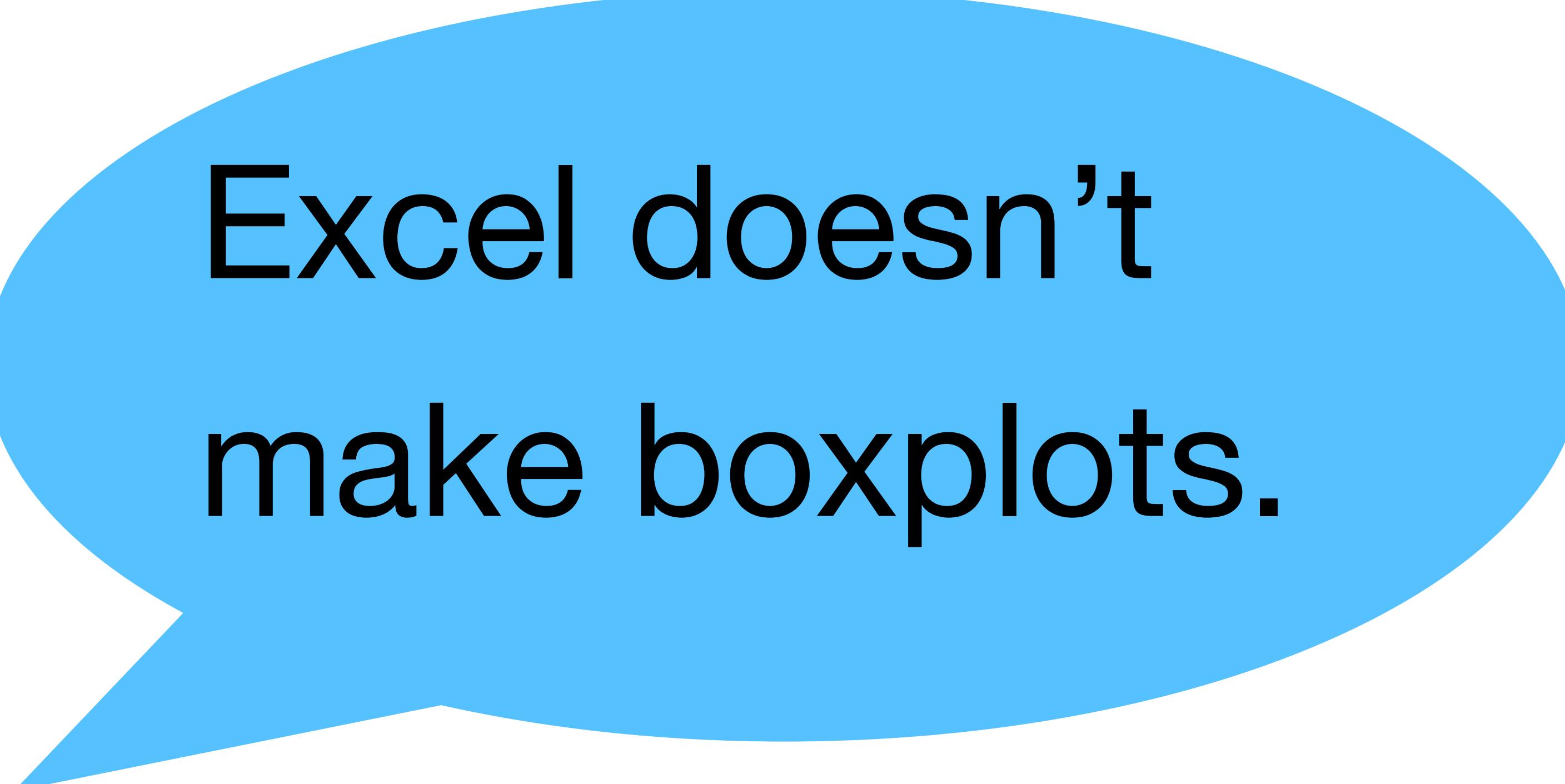
Crossroads of Biology

[Home](#)[Meetings](#)[Courses](#)[WELCOME](#)[INFO](#)[REGISTER](#)[TRAVEL](#)[ABSTRACTS](#)[SPONSORS](#)[PAYMENTS](#)[POLICIES](#)

Biological Data Science

November 4 - 7, 2020

Abstract Deadline (consideration for a talk): August 14, 2020



Excel doesn't
make boxplots.

Excel doesn't
make boxplots.

My data is not
normal, but I
want a p-value.

Excel doesn't
make boxplots.

My data is not
normal, but I
want a p-value.

Are there hidden
biases within my
data?

What do I need?

a reason

tools ready to be used

help, if getting stuck

Which computer?

Which computer?

for starting doesn't matter
(but 16GB memory is nice)

Which language?

Which language?

Maybe your field already defines it.

Otherwise...

Python

Python glue

Python glue

R
statistics,
political
science,
bio-
informatics

julia

numerical simulations

parts of
physics

R

statistics,
political
science,
bio-
informatics

Python
glue

julia

numerical simulations

parts of
physics

tableau
interactivity
business

R

statistics,
political
science,
bio-
informatics

Python
glue

julia

numerical simulations

parts of
physics

tableau

interactivity
business

Python
glue

R
statistics,
political
science,
bio-
informatics

matlab
engineering
images

Step 0 learn programming.

Step 0 learn programming.

courses (university, [datacamp.com](#))

Step 0 learn programming.

courses (university, [datacamp.com](#))

<https://github.com/amarallab/Introduction-to-Python-Programming-and-Data-Science>

Most important first step:

Get environment to work.

https://www.anaconda.com/distribution/

Products Why Anaconda? Solutions Resources Company Contact Us Download Search

Anaconda Individual Edition

The World's Most Popular Python/R Data Science Platform

Download

The open-source **Anaconda Individual Edition** (formerly **Anaconda Distribution**) is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 19 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling *individual data scientists* to:

- Quickly download 7,500+ Python/R data science packages
- Manage libraries, dependencies, and environments with **Conda**
- Develop and train machine learning and deep learning models with **scikit-learn**, **TensorFlow**, and **Theano**
- Analyze data with scalability and performance with **Dask**,

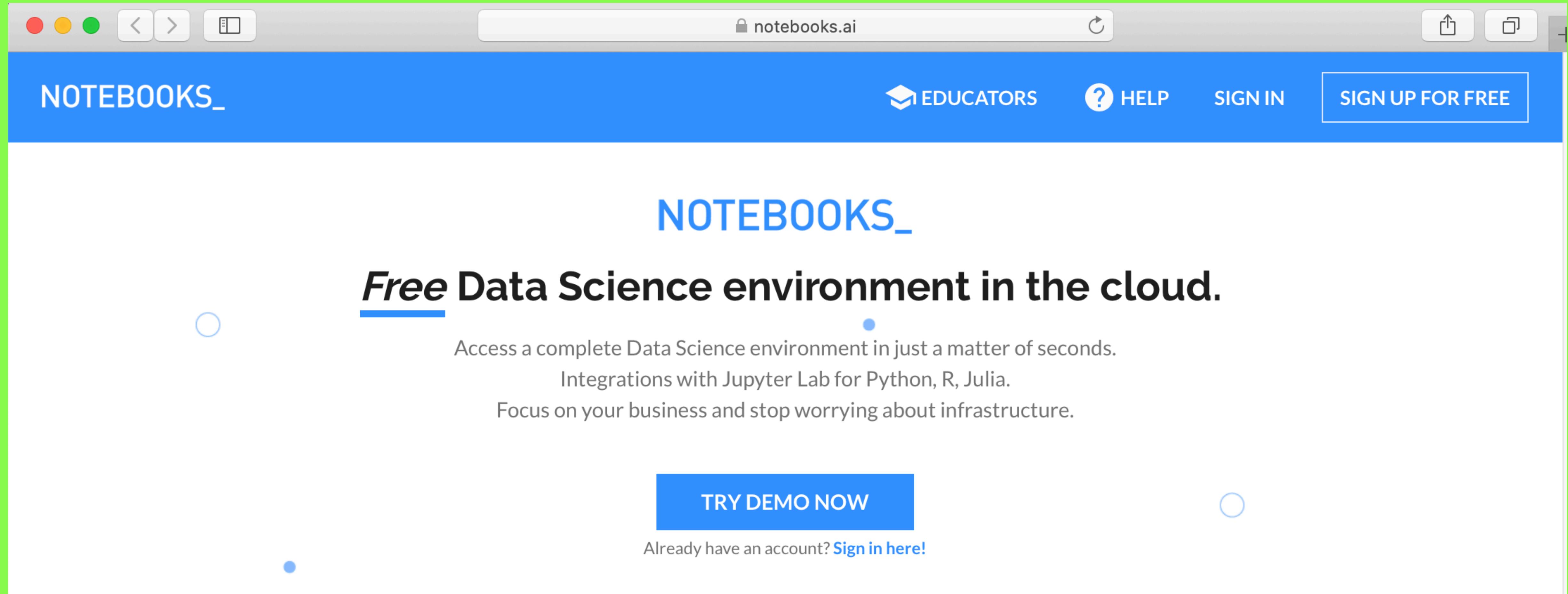
grid of 15 icons: jupyter, spyder, numpy, scipy, numba, pandas, dask, bokeh, holoviews, matplotlib, scikit-learn, h2o.ai, tensorflow, conda

download a distribution
that comes with
commonly used
“dependencies”

download latest version
(for python 3.7)

www.anaconda.com/distribution

Cloud provider, e.g.: notebooks.ai



The screenshot shows a web browser window for the website notebooks.ai. The page has a blue header bar with the text "NOTEBOOKS_" on the left, and "EDUCATORS", "HELP", "SIGN IN", and "SIGN UP FOR FREE" on the right. Below the header, the word "NOTEBOOKS_" is displayed in large blue letters. Underneath it, the text "Free Data Science environment in the cloud." is shown in bold black font, with "Free" underlined. To the left of this text is a small blue circle icon. Below the main title, there are three bullet points: "Access a complete Data Science environment in just a matter of seconds.", "Integrations with Jupyter Lab for Python, R, Julia.", and "Focus on your business and stop worrying about infrastructure.". To the right of the first bullet point is a small blue dot icon. At the bottom center is a blue button with the text "TRY DEMO NOW". Below the button, the text "Already have an account? [Sign in here!](#)" is visible. The background of the page features several small blue circular icons.

NOTEBOOKS_

EDUCATORS HELP SIGN IN SIGN UP FOR FREE

NOTEBOOKS_

Free Data Science environment in the cloud.

- Access a complete Data Science environment in just a matter of seconds.
- Integrations with Jupyter Lab for Python, R, Julia.
- Focus on your business and stop worrying about infrastructure.

TRY DEMO NOW

Already have an account? [Sign in here!](#)

Second most important initial step:

Getting some data.

Second most important initial step:

Getting some data.

e.g.: <https://www.kaggle.com/datasets>

Second most important initial step:

Getting some data.

e.g.: <https://www.kaggle.com/datasets>

ATTENTION: 80% of data science is data cleaning,
but example datasets often focus on the other 20%.

Second most important initial step:

Getting some data.

e.g.: <https://www.kaggle.com/datasets>

ATTENTION: 80% of data science is data cleaning,
but example datasets often focus on the other 20%.

Own research data!

What do I need?

a reason

tools ready to be used

help, if getting stuck

```
1 numbers_from_1_to_10 = np.arange(1, 11)
```

copy/paste
to google

```
1 numbers_from_1_to_10 = np.arange(1, 11)
```

```
NameError  
call last)
```

```
<ipython-input-1-b82cbf274997> in <module>  
----> 1 numbers_from_1_to_10 = np.arange(1, 11)
```

```
Traceback (most recent
```

```
NameError: name 'np' is not defined
```

copy/paste
to google

```
1 numbers_from_1_to_10 = np.arange(1, 11)
```

```
NameError  
call last)  
<ipython-input-1-b82cbf274997> in <module>  
----> 1 numbers_from_1_to_10 = np.arange(1, 11)
```

NameError: name 'np' is not defined

copy/paste
to google



All Videos Images Shopping News More Settings Tools

About 83,600 results (0.49 seconds)

stackoverflow.com › questions › error-nameerror-name...

[Error NameError: name 'np' is not defined - Stack Overflow](#)

Oct 22, 2018 - This imports the package numpy , and everything inside of that package.

However, numpy does **not** contain a module called np . The typical practice for numpy is to instead do import numpy as np. <https://stackoverflow.com/questions/52921955/error-nameerror-name-np-is-not-defined/52921990#52921990>.

<http://stackoverflow.com/>

coding

<http://datascience.stackexchange.com/>

data science

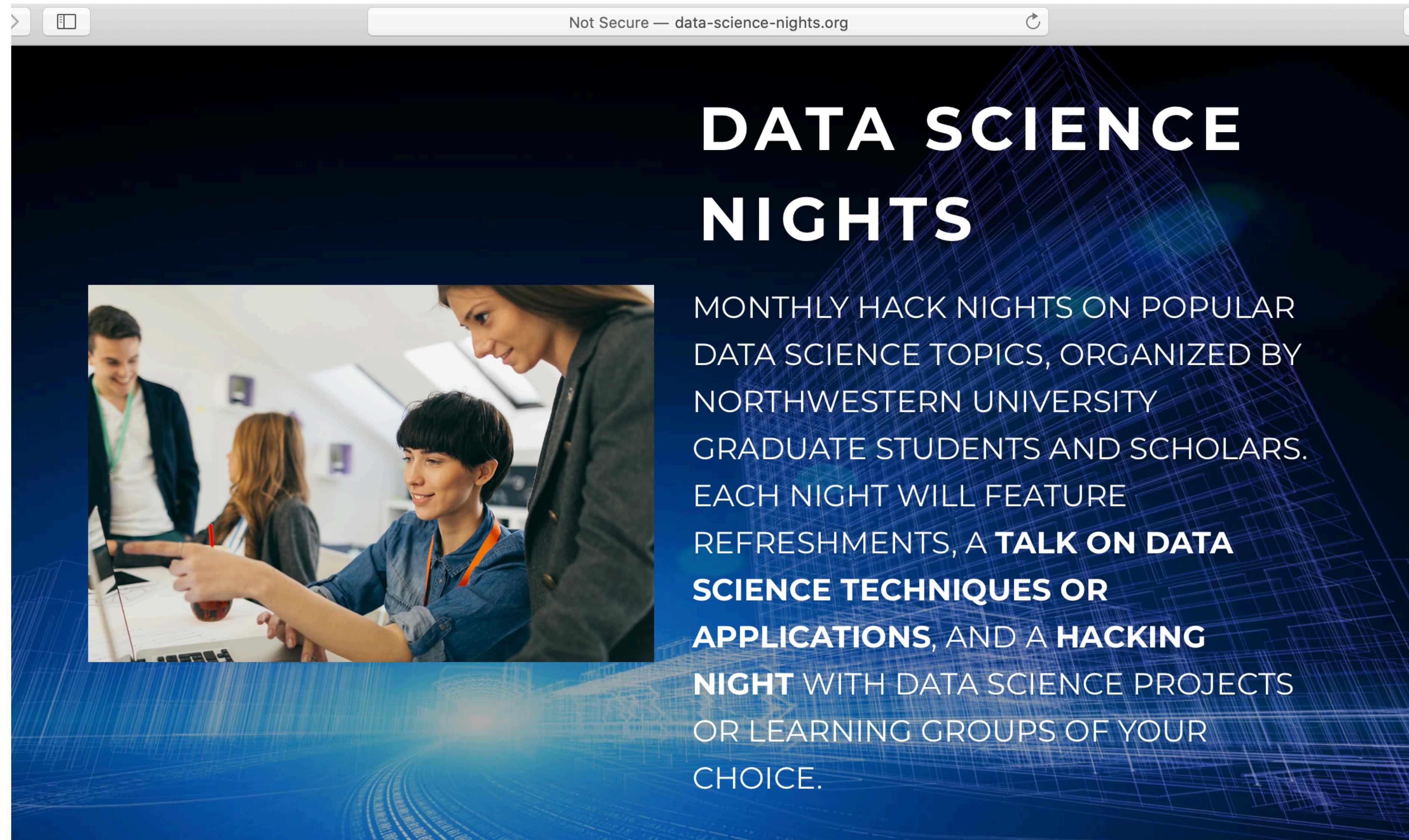
<http://biostars.org/>

bioinformatics

Universities

e.g.: free ebooks, consulting
services, courses

meetups

A screenshot of a web browser showing the homepage of data-science-nights.org. The page has a dark blue background with a faint wireframe grid. At the top, it says "Not Secure — data-science-nights.org". The main title "DATA SCIENCE NIGHTS" is in large white capital letters. Below the title is a photograph of four people at a table, looking at something together. To the right of the photo is a block of text describing the monthly hack nights. The URL "http://www.data-science-nights.org" is displayed at the bottom of the page.

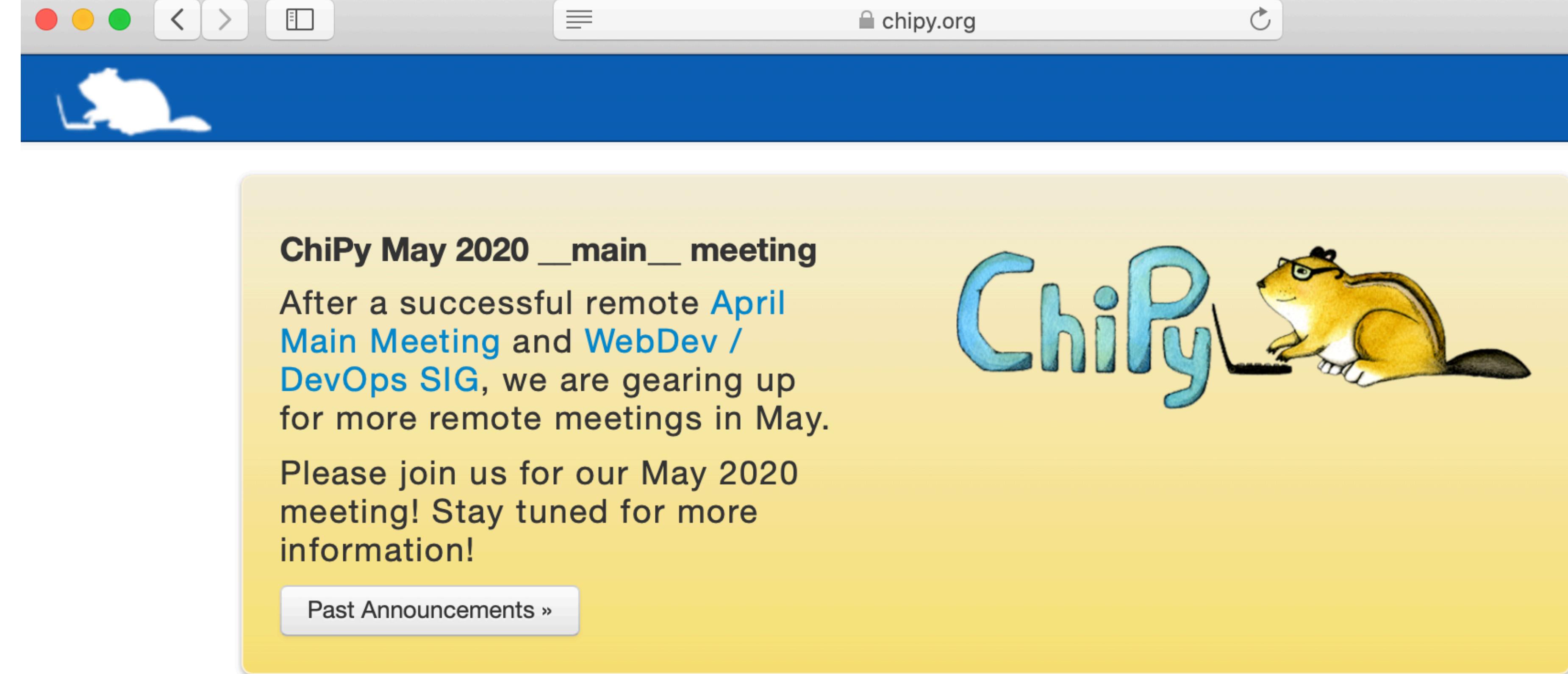
DATA SCIENCE NIGHTS



MONTHLY HACK NIGHTS ON POPULAR DATA SCIENCE TOPICS, ORGANIZED BY NORTHWESTERN UNIVERSITY GRADUATE STUDENTS AND SCHOLARS. EACH NIGHT WILL FEATURE REFRESHMENTS, A **TALK ON DATA SCIENCE TECHNIQUES OR APPLICATIONS**, AND A **HACKING NIGHT** WITH DATA SCIENCE PROJECTS OR LEARNING GROUPS OF YOUR CHOICE.

<http://www.data-science-nights.org>

meetups



The screenshot shows a web browser window with the URL <http://www.chipy.org> in the address bar. The page content is as follows:

ChiPy May 2020 __main__ meeting
After a successful remote [April Main Meeting](#) and [WebDev / DevOps SIG](#), we are gearing up for more remote meetings in May.
Please join us for our May 2020 meeting! Stay tuned for more information!

[Past Announcements »](#)



<http://www.chipy.org>

meetups

The screenshot shows a web browser window with the URL chipy.org in the address bar. The page features a blue header with a white chipmunk icon. Below the header is a yellow announcement box containing text about the May 2020 meeting and a link to past announcements. To the right of the text is a large blue "ChiPy" logo with a cartoon chipmunk character.

ChiPy May 2020 __main__ meeting

After a successful remote [April Main Meeting](#) and [WebDev / DevOps SIG](#), we are gearing up for more remote meetings in May.

Please join us for our May 2020 meeting! Stay tuned for more information!

[Past Announcements »](#)

ChiPy

<http://www.chipy.org>

<https://chipymentor.org>

What do I need?

What can I ignore?

deep learning

deep learning

as needed

deep learning

as needed

big data

deep learning

as needed

big data

as needed

deep learning

as needed

big data

as needed

virtual environments

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

version control

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

version control

in ~1 month

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

version control

in ~1 month

recommendation: use GitHub Desktop

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

version control

in ~1 month

recommendation: use GitHub Desktop

portfolio

deep learning

as needed

big data

as needed

virtual environments

in ~3 months

version control

in ~1 month

recommendation: use GitHub Desktop

portfolio

depends on reason
to learn data science

first steps:

first steps:

get working environment

first steps:

get working environment

get some data

first steps:

get working environment →

play and
learn

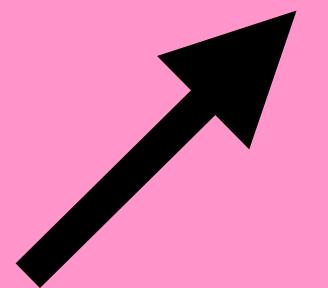
get some data

first steps:

get working environment →

play and
learn

get some data



ignore as much as possible

Bonus: install and start an environment