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

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.

Are there hidden biases within my data?

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

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

Python

glue

R

statistics,
political
science,
bio-
informatics

julia

numerical simulations
parts of
physics

R

statistics,
political
science,
bio-
informatics

Python

glue

tableau

interactivity
business

julia

numerical simulations
parts of
physics

R

statistics,
political
science,
bio-
informatics

Python

glue

tableau

interactivity
business

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/

ANACONDA.

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 (formally 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:

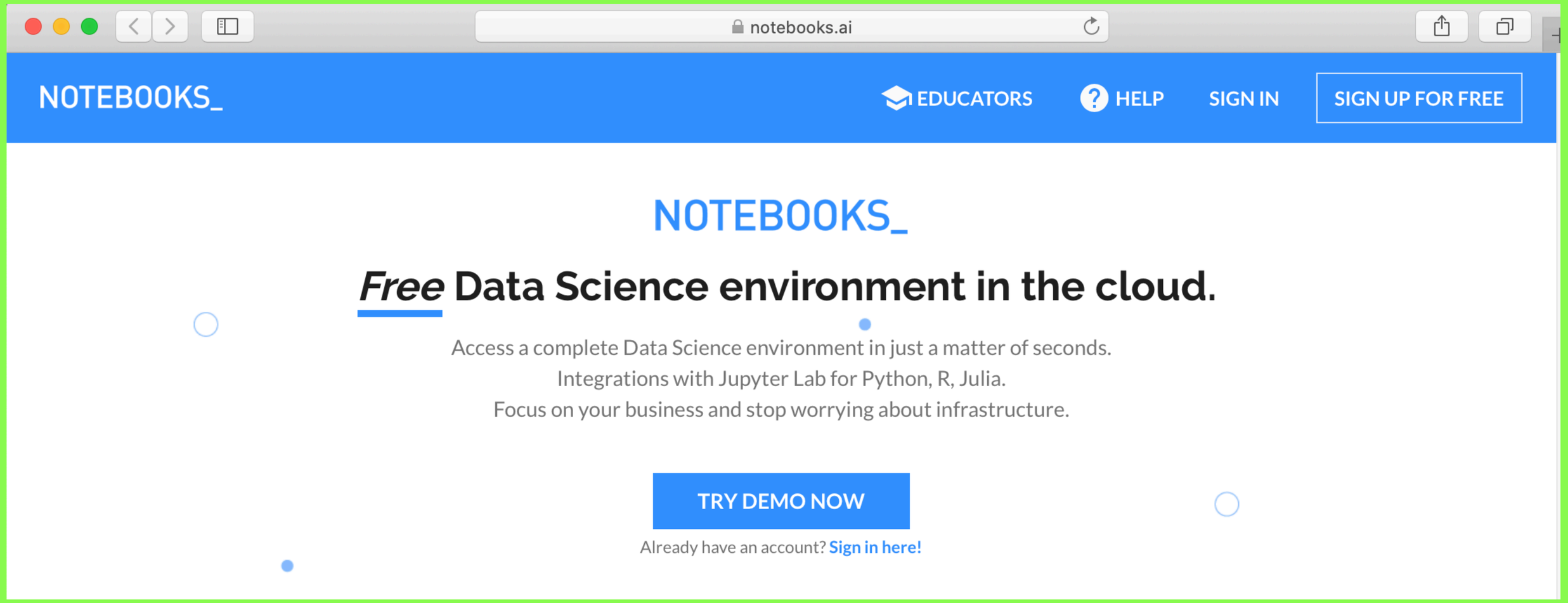
- Quickly download 7,500+ Python/R data science packages
- Manage libraries, dependencies, and environments with Conda
- Develop and train machine learning and deep learning models with scikit-learn, TensorFlow, and Theano
- Analyze data with scalability and performance with Dask,

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



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

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

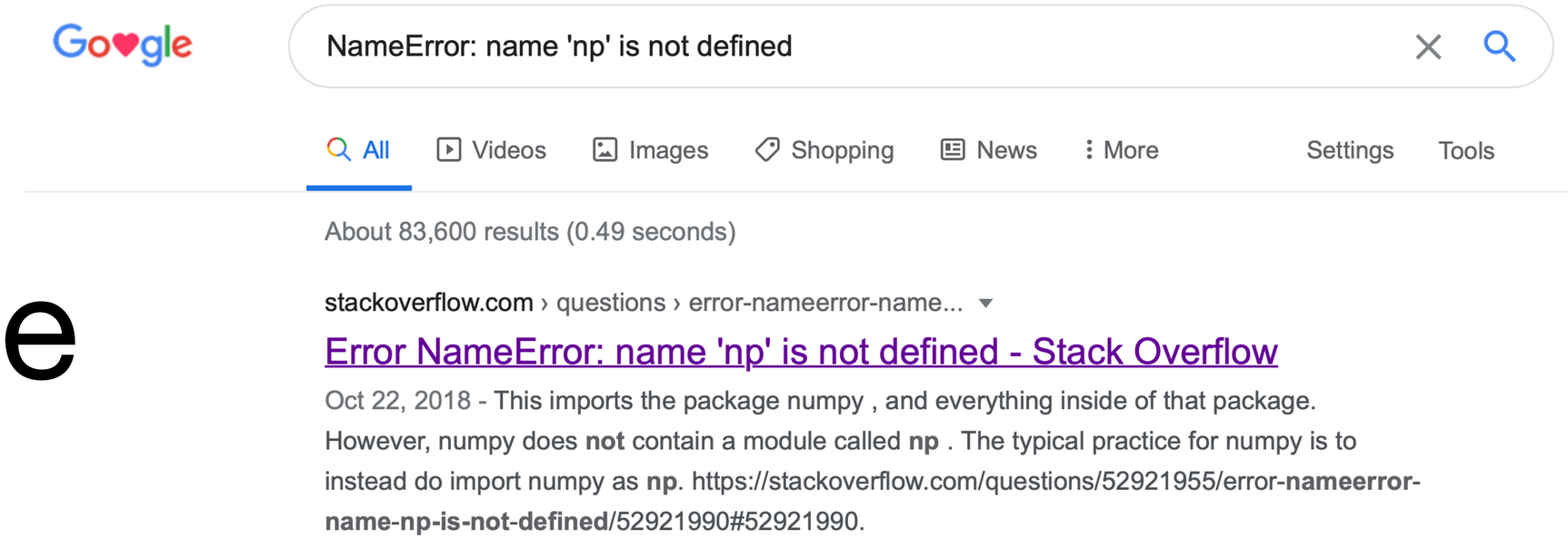
copy/paste to google

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent
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

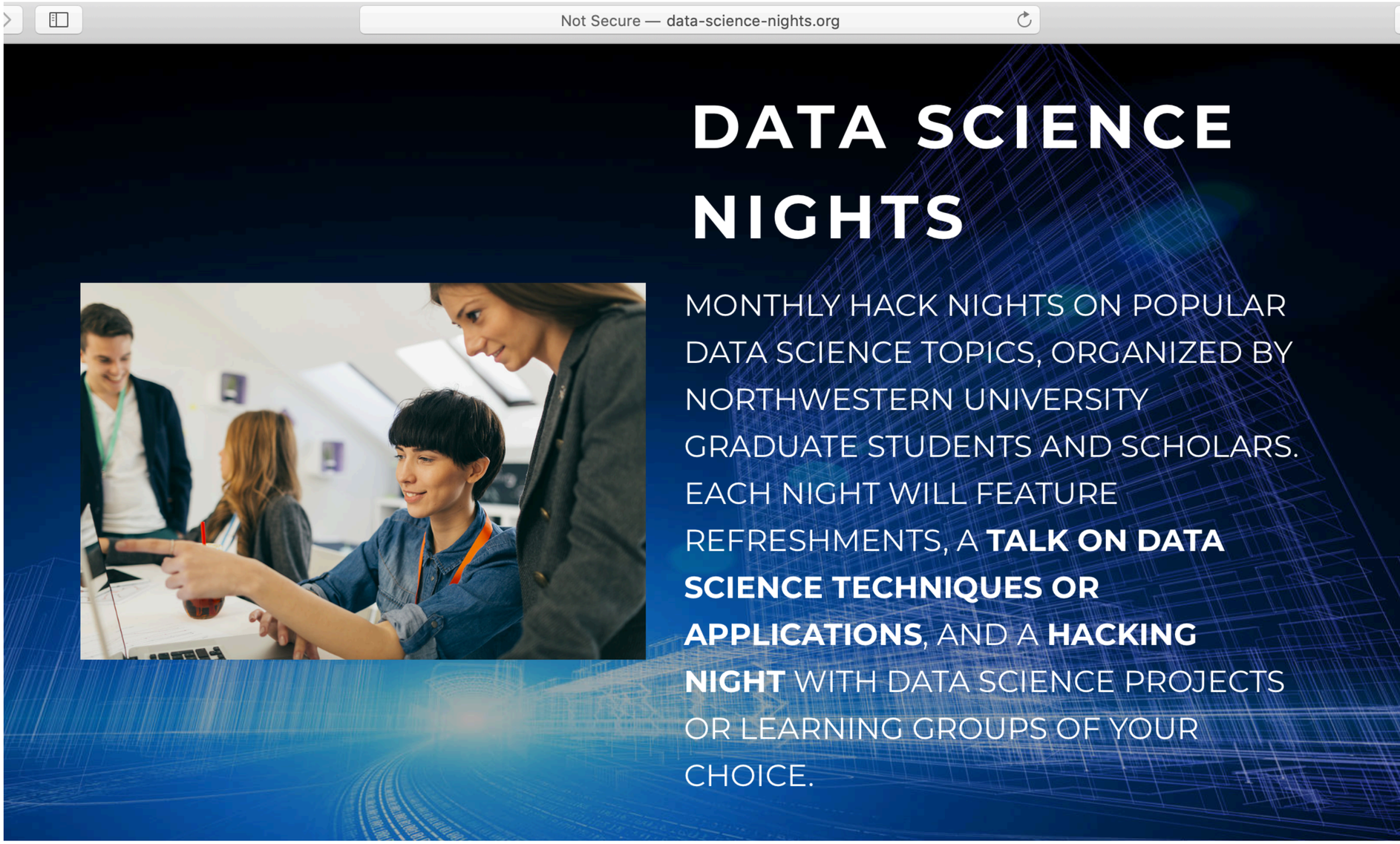


http://stackoverflow.com/	coding
http://datascience.stackexchange.com/	data science
http://biostars.org/	bioinformatics

Universities

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

meetups



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

meetups

http://www.chipy.org

meetups

http://www.chipy.org

https://chipymentor.org

What do I need?

What can I ignore?

deep learning

deep learning

as needed

deep learning

big data

as needed

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