

# COMM054: Data Science Principles & Practices



## Introduction to the Data Science Practices

Dr. Manal Helal

10 October 2019

# What Programming Language Do You Know?

[Back](#)

When poll is active, respond at **PollEv.com/mhelal**  
Text **MHELAL** to **07480 781235** once to join

**What Programming Language Do You Know?**

No responses received yet. They will appear here...

**Show responses** 

**Visual settings** 

**Activate** 

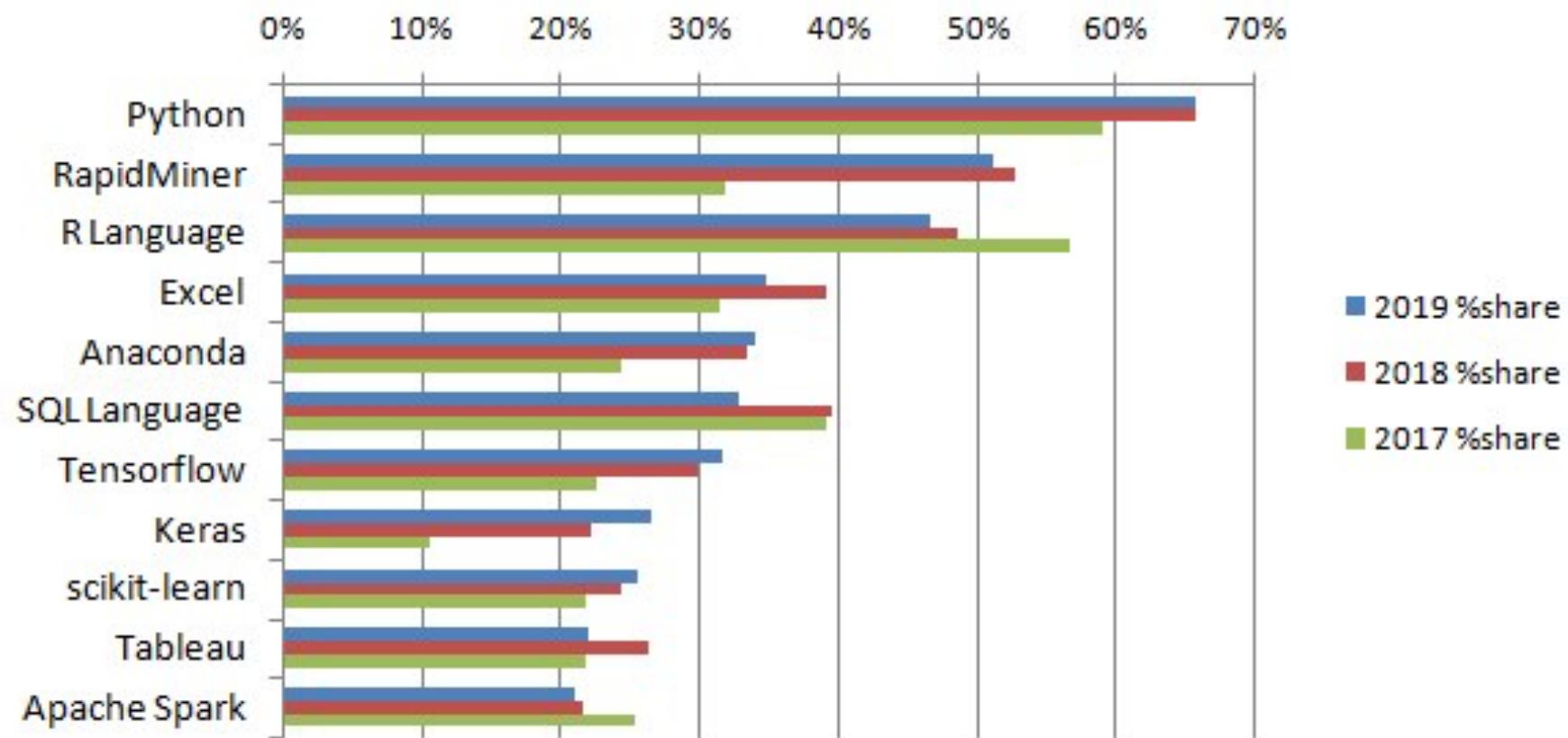
**Lock** 

**Clear responses** 

**Fullscreen** 

[Logout](#)

## Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



# Python

- Why Python?
  - *It's easy to learn*
    - *Now the language of choice for 8 of 10 top US computer science programs (Philip Guo, CACM)*
- *Full featured*
  - *Not just a statistics language, but has full capabilities for data acquisition, cleaning, databases, high performance computing, and more*
- *Strong Data Science Libraries*
  - *The SciPy Ecosystem*

# Planet Python

The screenshot shows a blog post titled "Real Python" with the subtitle "Emacs: The Best Python Editor?". The post discusses the challenges of choosing a code editor for Python development and highlights Emacs as a feature-rich option. It includes a list of steps for setting up Emacs for Python development and notes that the tutorial is for GNU Emacs 25 or later. There is also an "Updates" section with two entries: one from October 9, 2019, and another from November 3, 2015. At the bottom, there's a link to download code samples and a "Screenshot" button.

python™

RSS feed  
Titles Only  
Powered by Planet!

Other Python Planets  
Python Summer of Code  
Planet Python Francophone  
Planet Python Argentina  
Planet Python Japan  
Planet Python Brasil  
Planet Python Indonesia  
Planet Python Poland

Python Libraries  
PySoy  
SciPy  
SymPy  
Twisted  
Python/Web Planets  
CherryPy  
Django Community  
Plone  
TurboGears

Other Languages  
Haskell  
Lisp  
Parrot  
Perl  
Ruby

Databases  
MySQL  
PostgreSQL

Subscriptions

Planet Python

Last update: October 09, 2019 09:48 PM UTC

October 09, 2019

**Real Python**

**Emacs: The Best Python Editor?**

Finding the right code editor for Python development can be tricky. Many developers explore numerous editors as they grow and learn. To choose the right code editor, you have to start by knowing which features are important to you. Then, you can try to find editors that have those features. One of the most feature-rich editors available is [Emacs](#).

Emacs started in the mid-1970s as a set of macro extensions for a different code editor. It was adopted into the [GNU project](#) by Richard Stallman in the early 1980s, and GNU Emacs has been continuously maintained and developed ever since. To this day, GNU Emacs and the XEmacs variant are available on every major platform, and GNU Emacs continues to be a combatant in the [Editor Wars](#).

In this tutorial, you'll learn about using Emacs for Python development, including how to:

- Install Emacs on your selected platform
- Set up an Emacs initialization file to configure Emacs
- Build a basic Python configuration for Emacs
- Write Python code to explore Emacs capabilities
- Run and Test Python code in the Emacs environment
- Debug Python code using integrated Emacs tools
- Add source control functionality using Git

For this tutorial, you'll use GNU Emacs 25 or later, although most of the techniques shown will work on older versions (and XEmacs) as well. You should have some experience developing in Python, and your machine should have a Python distribution already [installed](#) and ready to go.

**Updates:**

- 10/09/2019: Major update adding new code samples, updated package availability and info, basic tutorial, Jupyter walk-through, debugging walk-through, testing walk-through, and updated visuals.
- 11/03/2015: Initial tutorial published.

You can download all the files referenced in this tutorial at the link below:

Download Code: [Click here to download the code you'll use](#) to learn about Emacs for Python in this tutorial.

Installation and Basics

Screenshot

- <http://planetpython.org/>
- Excellent blog aggregator for python related news
- Significant number of data science and python tutorials are posted
- Great blend of applied beginner and higher level python postings

# What are Python Notebooks?

- A notebook combines the functionality of
  - a word processor — handles formatted text
  - a "shell" or "kernel" — executes statements in a programming language and includes output inline
  - a rendering engine — renders : HTML in addition to plain text

# Notebook benefits

- A single document that combines explanations with executable code and its output — an ideal way to provide:
  - reproducible research results
  - documentation of processes
  - instructions
  - tutorials and training materials of all shapes and sizes
- A digital learning environment for computational thinking

# What is Jupyter notebook?

- First notebook: Mathematica 1.0 in '88
- Jupyter notebook is a part of Project Jupyter, a nonprofit organization that:
  - develops open-source software,
  - provides standards, and services for interactive computing across dozens of programming languages
- Beginning with Julia, Python, R, and now over 70 languages.

# Ways to use Jupyter notebooks

- On your own computer after installing a Jupyter notebook server (It is installed in the labs: `~/.local/bin/jupyter`)
- With an online notebook server like [cocalc.com](https://cocalc.com)
- Save notebook with output and use a notebook viewer
- Export to HTML, PDF, LATEX, etc.

# Static notebook viewers

- Notebooks are stored as .ipynb files
  - .ipynb files are in JSON format
  - each code cell may include output from the last execution
- You can share an .ipynb file
  - anyone with a local notebook server can view it
  - ... but of course cannot execute anything new
- Many online systems have viewers
  - GitHub's file previewer
  - Project Jupyter's [nbviewer.jupyter.org](http://nbviewer.jupyter.org)

# Other Interesting Tools

- Wolfram Alpha computational knowledge engine, which processes natural language-like queries through a mix of algorithms and pre-digested data sources.
- Matlab is for processing matrices, with many linear algebra, statistical and machine learning toolboxes. It is proprietary but GNU Octave is an open-source alternative.
- R is open source packages for statisticians. Many of its functions are available in Python, and you can call R library functions from Python if not available.
- Excel and other spreadsheet programs include a variety of functions for exploratory data analysis and data manipulations that is worth searching deep for before writing your own code.
- Google Books NGram shows the statistics of phrases usage as it evolves over time, with links to books mentioning these phrases. This enables any data scientist to check the context of using phrases and what they mean and how it evolves.
- Google Trends as well gives you the popularity of world top searches by region and language over time.
- For Big Data Analytics, Hadoop and Spark are the famous parallel processing systems using Sparc, Java and C++.

< Back

When poll is active, respond at **PollEv.com/mhelal**

Text **MHELAL** to **07480 781235** once to join

Visual settings 

Activate 

Show responses 

Lock 

Clear responses 

Fullscreen 

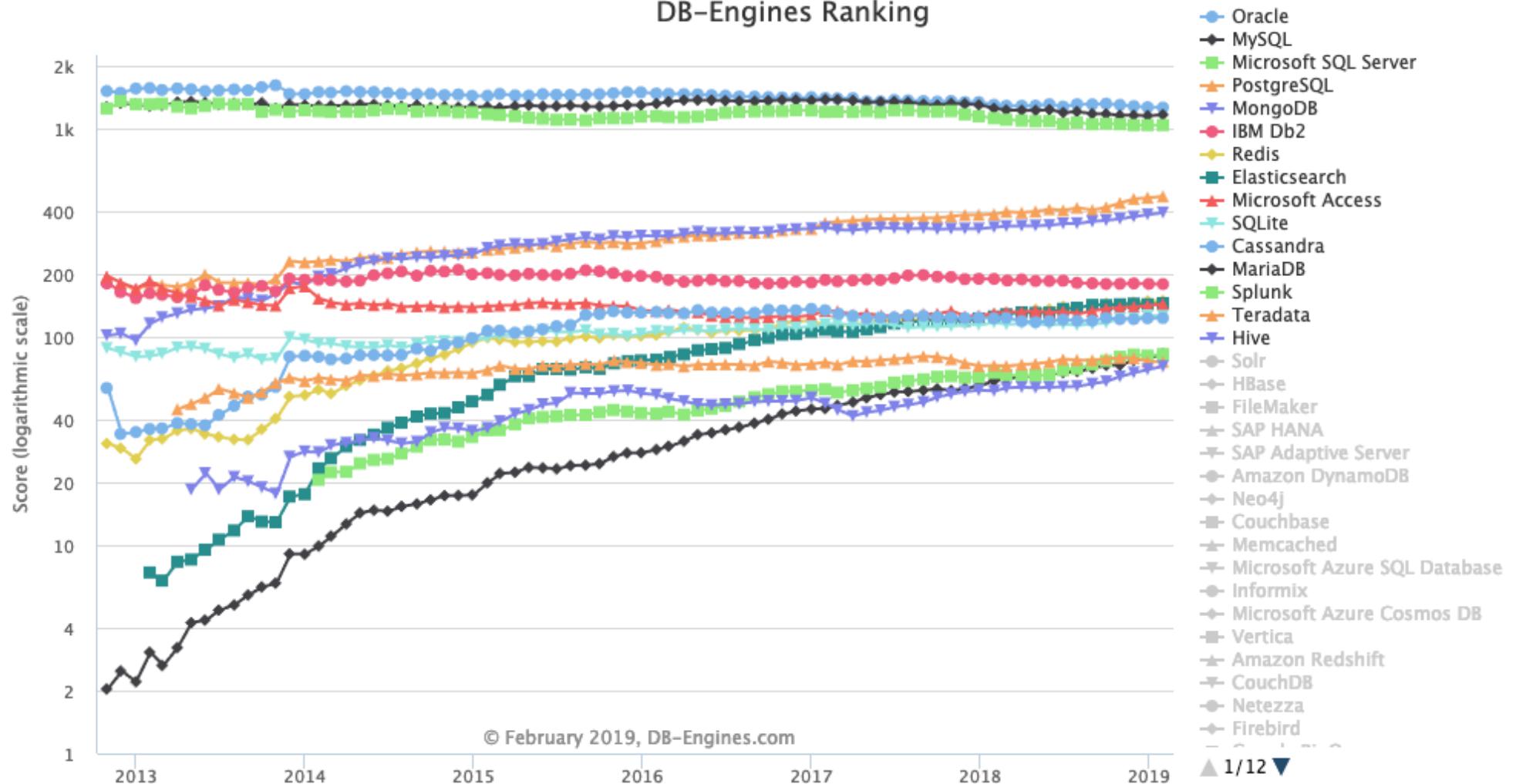
## What RDBMS do you know?



No responses received yet. They will appear here...

Logout

## DB-Engines Ranking



# Data Formats

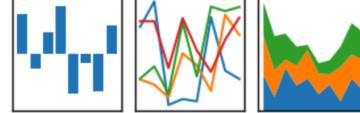
- Structured Data:
  - RDBMs servers and enquired using SQL (Structured Query Languages).
  - CSV (comma separated value) files: covered in the “lab1\_PyIntro.ipynb”
  - XML (eXtensible Markup Language): covered in the “lab1\_db.ipynb”
  - JSON (JavaScript Object Notation): covered in the “lab1\_db.ipynb”
- Unstructured Data:
  - Natural Language Text: Papers, Books, Journals, Blogs, websites, Facebook, Twitter, ... and so forth.
  - Multimedia Files: Images, Audio and Video.

# Pandas

- Created in 2008 by Wes McKinney
- Open source New BSD license
- 100 different contributors

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



[home](#) // [about](#) // [get pandas](#) // [documentation](#) // [community](#) // [talks](#) // [donate](#)

## Python Data Analysis Library

*pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](#) programming language.

*pandas* is a [NumFOCUS](#) sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to [donate](#) to the project.



### v0.25.1 Final (August 22, 2019)

This is a minor bug-fix release in the 0.25.x series and includes some regression fixes and bug fixes. We recommend that all users upgrade to this version.

See the [full whatsnew](#) for a list of all the changes.

#### VERSIONS

**Release**  
0.25.1 - August 2019  
[download](#) // [docs](#) // [pdf](#)

**Development**  
0.26.0 - September 2019  
[github](#) // [docs](#)

**Previous Releases**  
0.25.0 - [download](#) // [docs](#) // [pdf](#)  
0.24.2 - [download](#) // [docs](#) // [pdf](#)  
0.24.1 - [download](#) // [docs](#) // [pdf](#)  
0.24.0 - [download](#) // [docs](#) // [pdf](#)  
0.23.4 - [download](#) // [docs](#) // [pdf](#)  
0.23.3 - [download](#) // [docs](#) // [pdf](#)  
0.23.2 - [download](#) // [docs](#) // [pdf](#)  
0.23.1 - [download](#) // [docs](#) // [pdf](#)  
0.23.0 - [download](#) // [docs](#) // [pdf](#)  
0.22.0 - [download](#) // [docs](#) // [pdf](#)  
0.21.1 - [download](#) // [docs](#) // [pdf](#)  
0.21.0 - [download](#) // [docs](#) // [pdf](#)  
0.20.3 - [download](#) // [docs](#) // [pdf](#)  
0.20.2 - [download](#) // [docs](#) // [pdf](#)  
0.20.1 - [Screenshot](#) // [ad](#) // [docs](#) // [pdf](#)  
0.20.0 - [Screenshot](#) // [ad](#) // [docs](#) // [pdf](#)

# The Series

	Animals	Name
Index		Values
0	Dog	
1	Bear	
2	Tiger	
3	Moose	
4	Giraffe	
5	Hippopotamus	
6	Mouse	

# The DataFrame



# Stack Overflow

- <https://stackoverflow.com/questions>
- Massive knowledge forum of python and pandas related content
- Free to join and participate in
- Heavily used by pandas developers instead of a mailing list

The screenshot shows the Stack Overflow homepage. The main area displays a list of questions under the heading "All Questions". There are 18,339,312 questions listed. The first question is titled "Update With Multiple Condition" and has 0 votes and 0 answers. The second question is titled "How can you use the actual regex value in an environment variable?" and has 0 votes and 0 answers. The third question is titled "Printing out the values of Array" and has 0 votes and 0 answers. The fourth question is titled "Outlier detection and removal in PySpark" and has 0 votes and 0 answers. The sidebar on the right includes links to "Blog", "Adding Static Code Analysis to Stack Overflow", "Lessons from Design School for Software Engineers", "Featured on Meta", "Unicorn Meta Zoo #9: How do we handle problem users?", "An apology to our community, and next steps", and "Threshold experiment results: closing, editing and reopening all become more...". At the bottom of the sidebar, there is a Microsoft Azure advertisement with the text "Build and develop apps with Azure. Free until you say otherwise." and a "Try Azure Free" button.

# Open Access Books

- Springer:
    - <https://link.springer.com/search/page/3?facet-content-type=%22Book%22&package=openaccess>
  - KDnuggets is an online platform on business analytics, big data, data mining, and data science.



Subscribe to KDnuggets News | [Twitter](#) | [Facebook](#) | [LinkedIn](#)

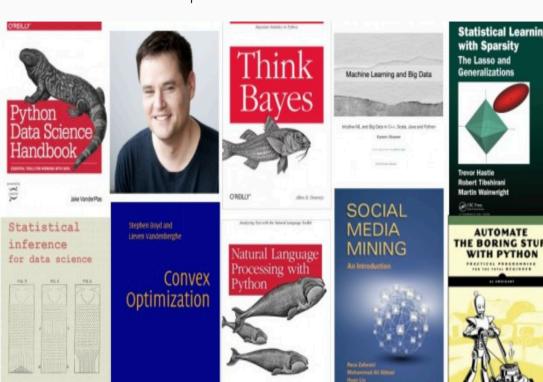
Contact

SOFTWARE | News/Blog | Top stories | Opinions | Tutorials | JOBS | Companies | Courses | Datasets | EDUCATION | Certificates | Meetings | Webinars

By Matthew Mayo, KDnuggets.

It's time for another collection of free machine learning and data science books to kick off your summer learning season. Because that's a thing. Right?

If, after reading this list, you find yourself wanting more free quality, curated books, [check the previous iteration of this series](#) or the related posts below.



By Matthew Mayo, KDnuggets.

It's time for another collection of free machine learning and data science books to kick off your summer learning season. Because that's a thing. Right?

If, after reading this list, you find yourself wanting more free quality, curated books, [check the previous iteration of this series](#) or the related posts below.



Help earthquake victims in Nepal.

[WATCH VIDEO >](#)

#data4good



SAS Brings AI and Analytics to the Cloud

---

MACHINE LEARNING VITAL SIGNS

Register Now

DOMINO MINER & KASCH



# SpringerLink

data science

New Search  

Home • Books A - Z • Journals A - Z • Videos • Librarians

Include Preview-Only content 

**60 Result(s) for 'data science' within Book**  

You are now only searching within the **Open Access** package  
STOP searching within this package

Refine Your Search

Content Type

Book 

Conference Proceedings	14
------------------------	----

Discipline

see all 

Computer Science	27
Education	6
Social Sciences	6
Earth Sciences	4
Environment	3

Subdiscipline

see all 

Information Systems Applications (incl. Internet)	11
Data Mining and Knowledge Discovery	7
Systems and Data Security	7

Sort By  Relevance  Newest First  Oldest First  Date Published  Page  1 of 3 

Book

## Fundamentals of Clinical Data Science

Dr. Pieter Kubben, Michel Dumontier... (2019)



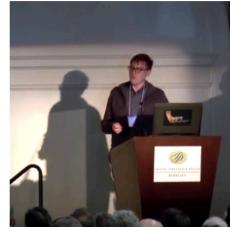
Book

## Data and Text Processing for Health and Life Sciences

Ph.D. Francisco M. Couto in *Advances in Experimental Medicine and Biology* (2019)



# Data Skeptic Podcast



- <http://dataskeptic.com/>
- Kyle Polich, created in 2014
- Covers data science more generally, including:
  - Mini educational lessons
  - Interviews
  - Trends
  - Shared community project (OpenHouse)

## A Skeptic's Perspective on AI

Last summer, I had the privilege of presenting for the 5th annual SkeptiCal Con 2018. The videos recently went live. I am very proud of this presentation and would appreciate you sharing it with anyone you think might enjoy it. Please check out my talk! [View More >](#)

May 21, 2018

## Notes on Uber's Experimentation Platform

I attended the Chicago AI & Data Science Conference 2018 this past weekend. I presented a talk, lead a panel, and had the opportunity to enjoy a number of interesting talks. The particular talk I took the most from was given by Jeremy Gu and Mandie Liu, both from Uber on their experimental platform. [View More >](#)

April 29, 2018

## Reinforcement Learning in the Real World

We've pulled together a few short videos that give interesting examples of robots that leverage reinforcement learning to do some physical task. [View More >](#)

