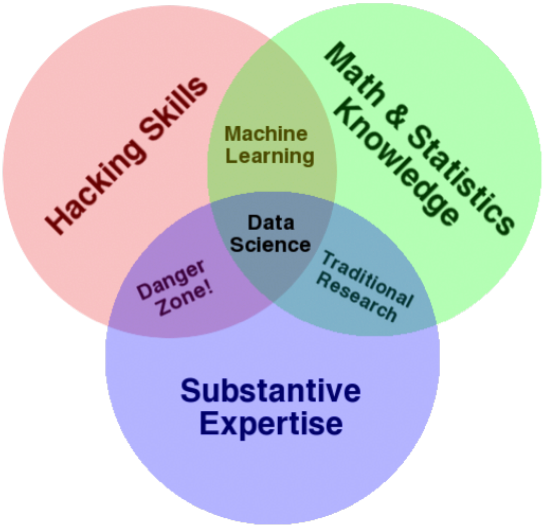


COMM054: Data Science Principles & Practices

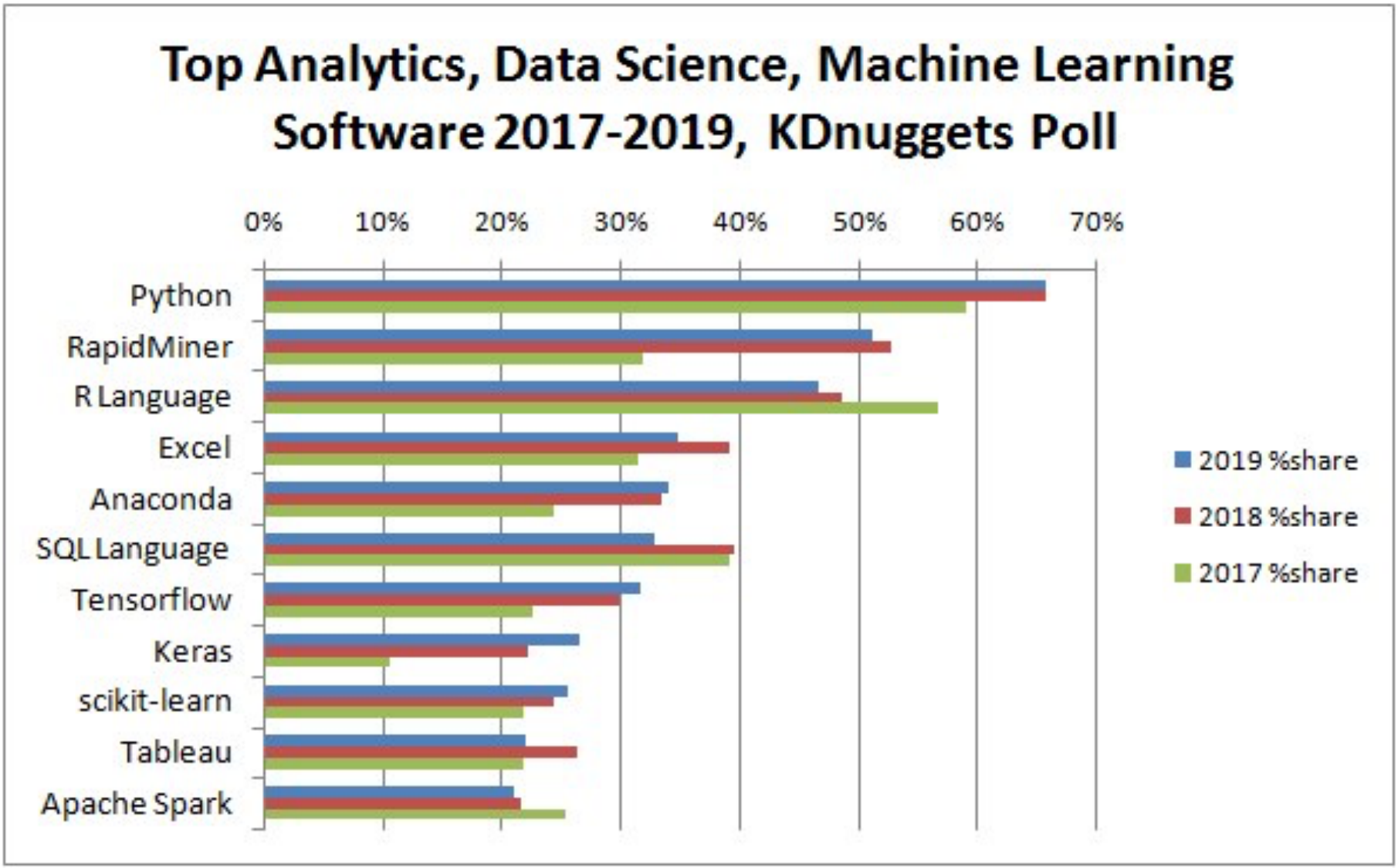
UNIVERSITY OF SURREY

Introduction to the Data Science Practices

Dr. Manal Helal

10 October 2019

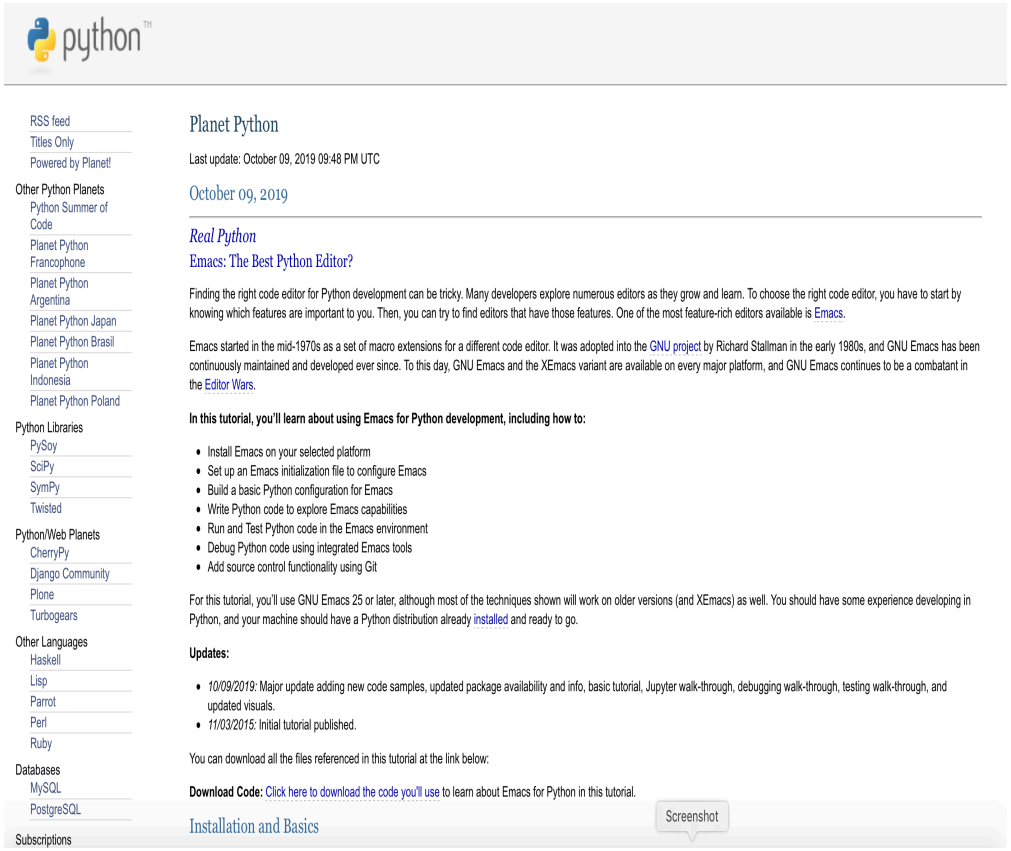
What Programming Language Do You Know?


Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R Language
Excel
Anaconda
SQL Language
Tensorflow
Keras
scikit-learn
Tableau
Apache Spark
2019 %share
2018 %share
2017 %share

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



- 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
- 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

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 propriety 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

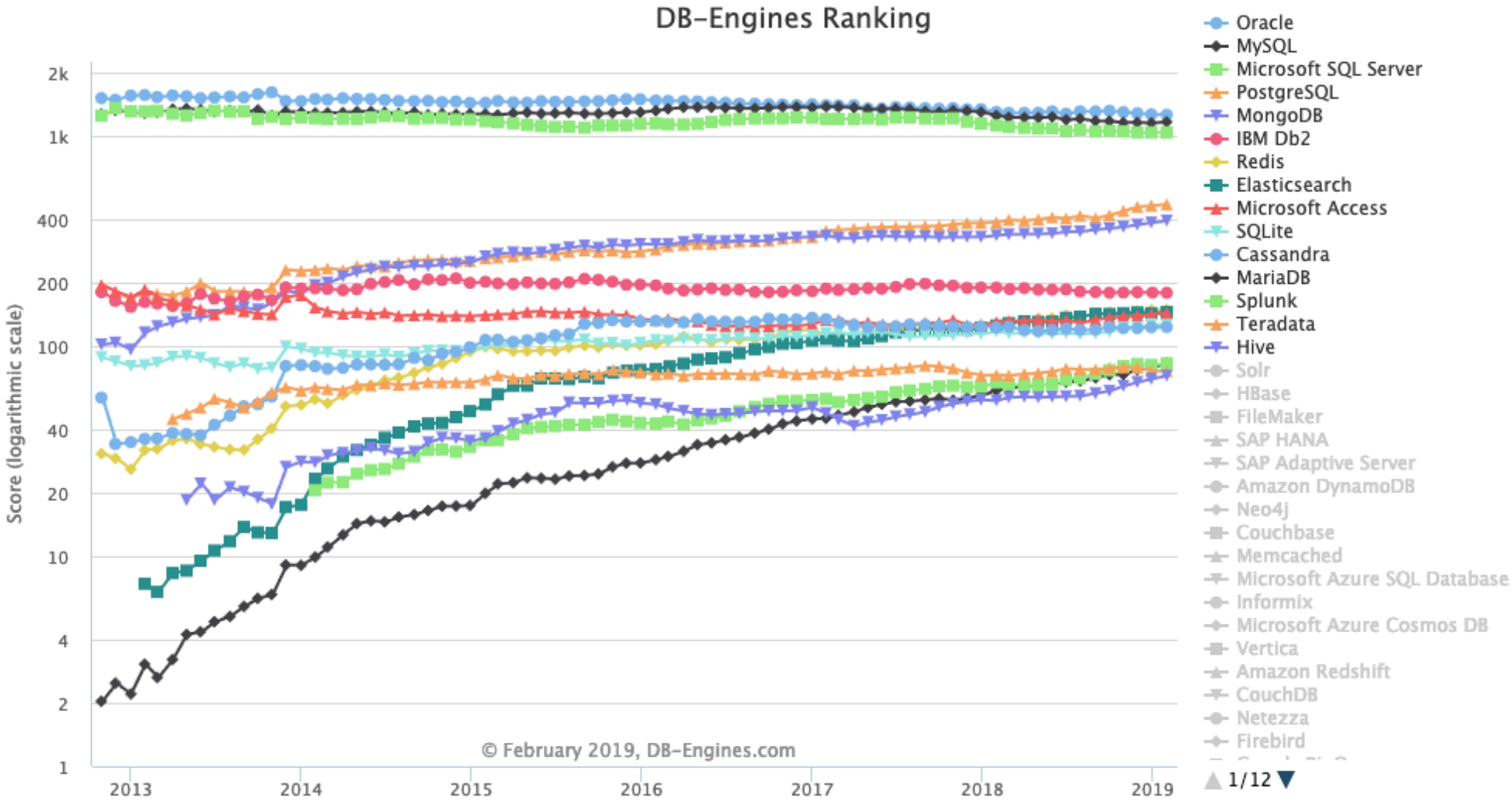
Clear responses

Fullscreen

What RDBMS do you know?

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

Logout

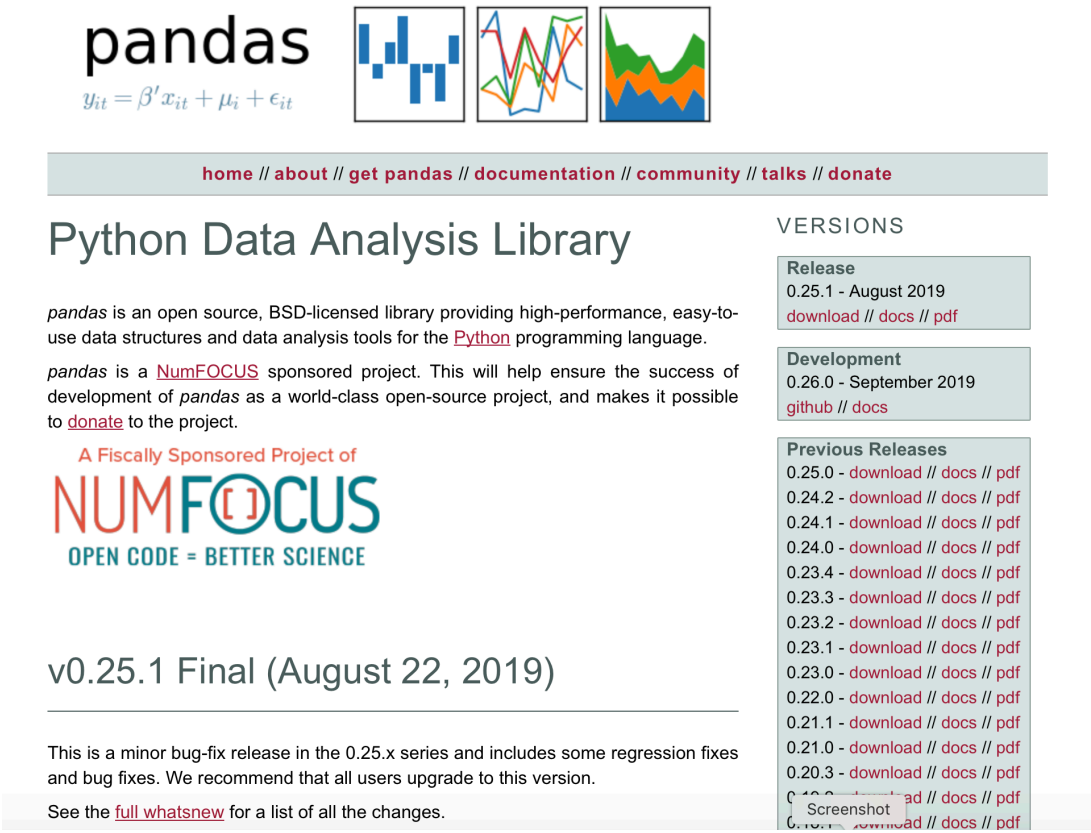

DB-Engines Ranking
Score (logarithmic scale)
2k
1k
400
200
100
40
20
10
4
2
1
2013
2014
2015
2016
2017
2018
2019
© February 2019, DB-Engines.com
Oracle
MySQL
Microsoft SQL Server
PostgreSQL
MongoDB
IBM Db2
Redis
Elasticsearch
Microsoft Access
SQLite
Cassandra
MariaDB
Splunk
Teradata
Hive
Solr
HBase
FileMaker
SAP HANA
SAP Adaptive Server
Amazon DynamoDB
Neo4j
Couchbase
Memcached
Microsoft Azure SQL Database
Informix
Microsoft Azure Cosmos DB
Vertica
Amazon Redshift
CouchDB
Netezza
Firebird
1/12

Data Formats

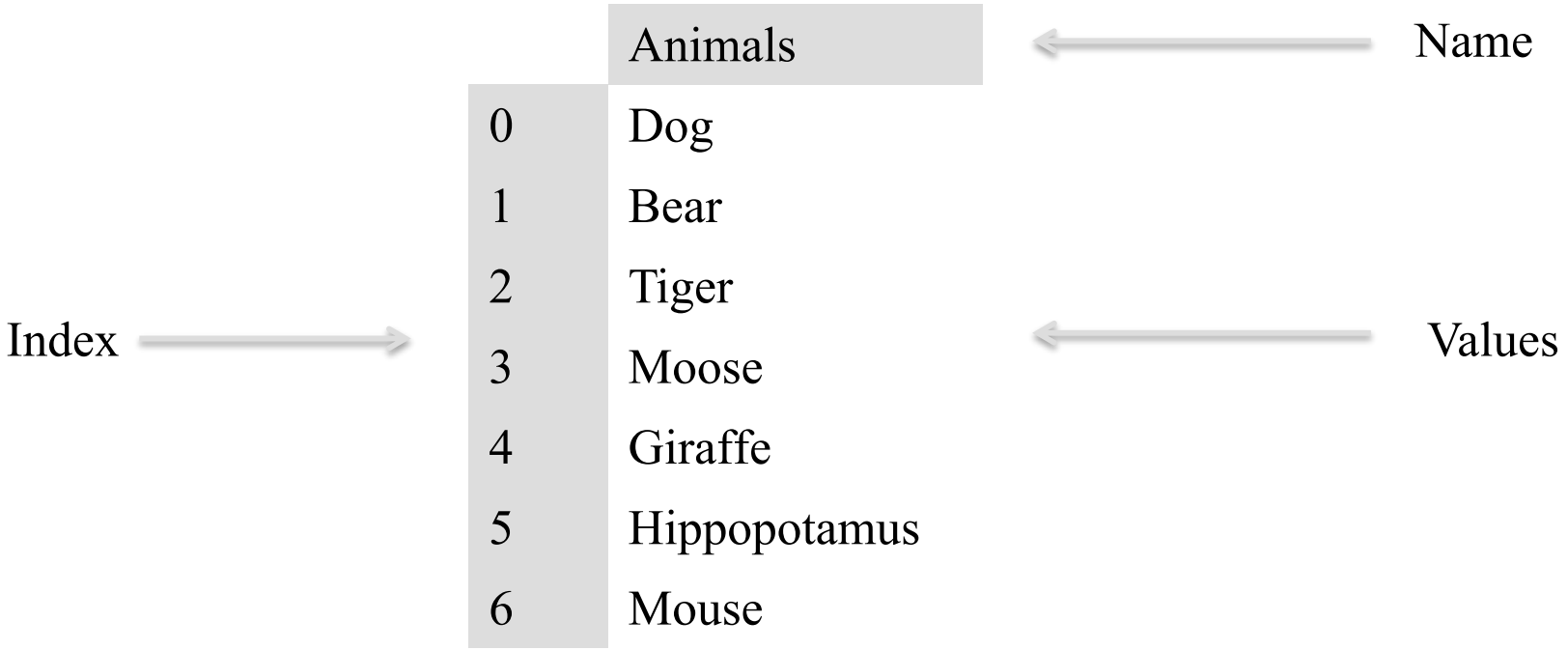
- Structured Data:
  - RBDMs servers and enquired using SQL (Structured Query Languages).
  - CSV (comma separated value) files: covered in the “lab1_PyIntro.ipynb”
  - XML (eXtensible Markup Language): covered in the “lab1_db.ipynb”
  - JSON (JavaScript Object Notation): covered in the “lab1_db.ipynb”
- Unstructured Data:
  - Natural Language Text: Papers, Books, Journals, Blogs, websites, Facebook, Twitter, … and so forth.
  - Multimedia Files: Images, Audio and Video.

Pandas

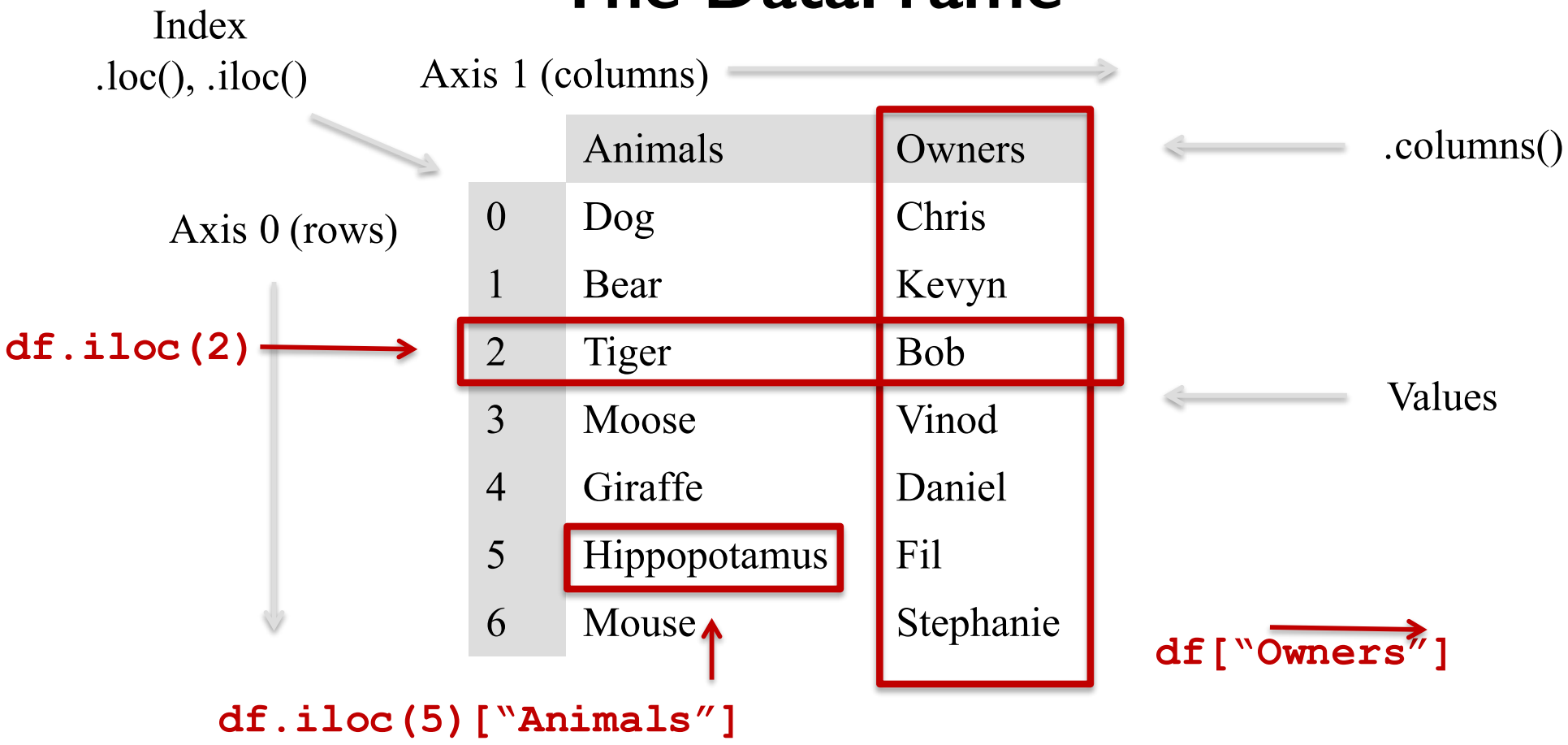
- Created in 2008 by Wes McKinney
- Open source New BSD license
- 100 different contributors



The Series


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

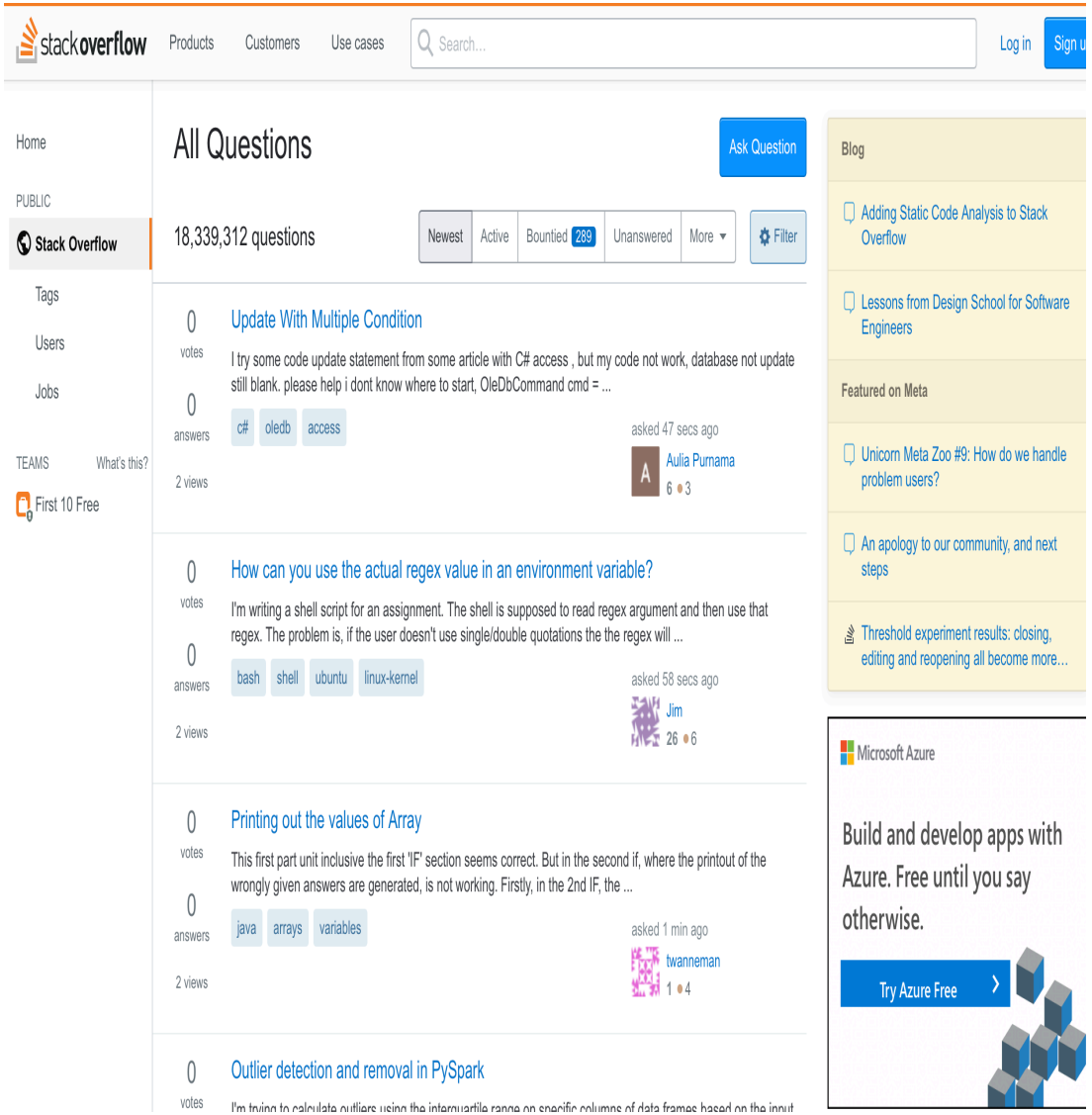
The DataFrame



	Animals	Owners
0	Dog	Chris
1	Bear	Kevyn
2	Tiger	Bob
3	Moose	Vinod
4	Giraffe	Daniel
5	Hippopotamus	Fil
6	Mouse	Stephanie

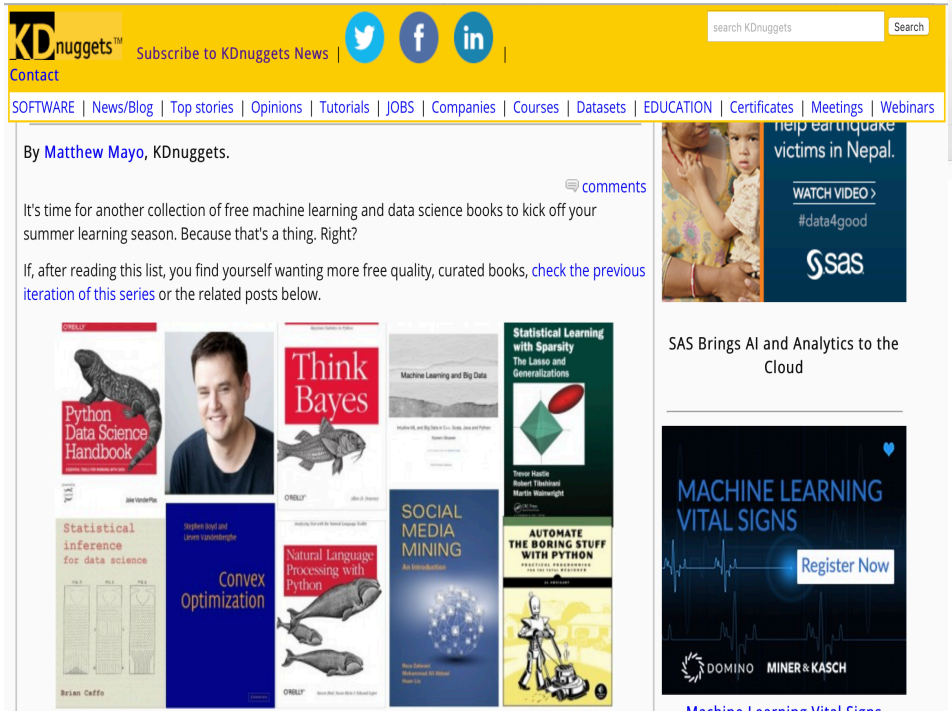
Stack Overflow

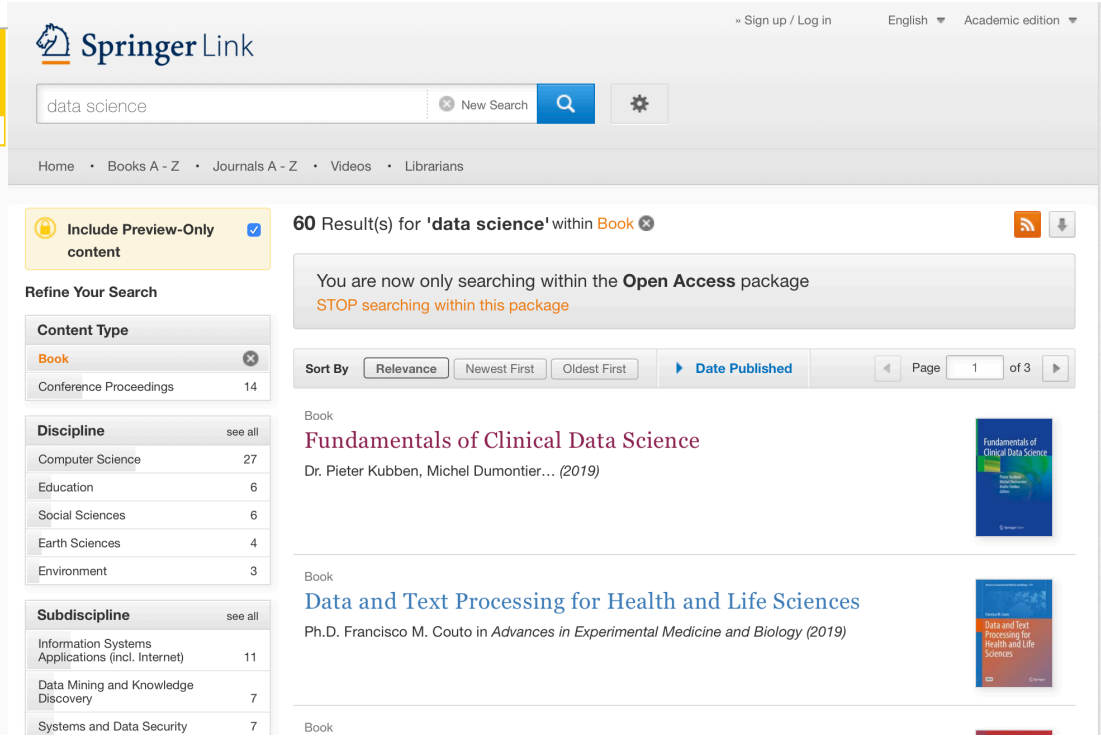
- https://stackoverflow.com/questions
- Massive knowledge forum of python and pandas related content
- Free to join and participate in
- Heavily used by pandas developers instead of a mailing list



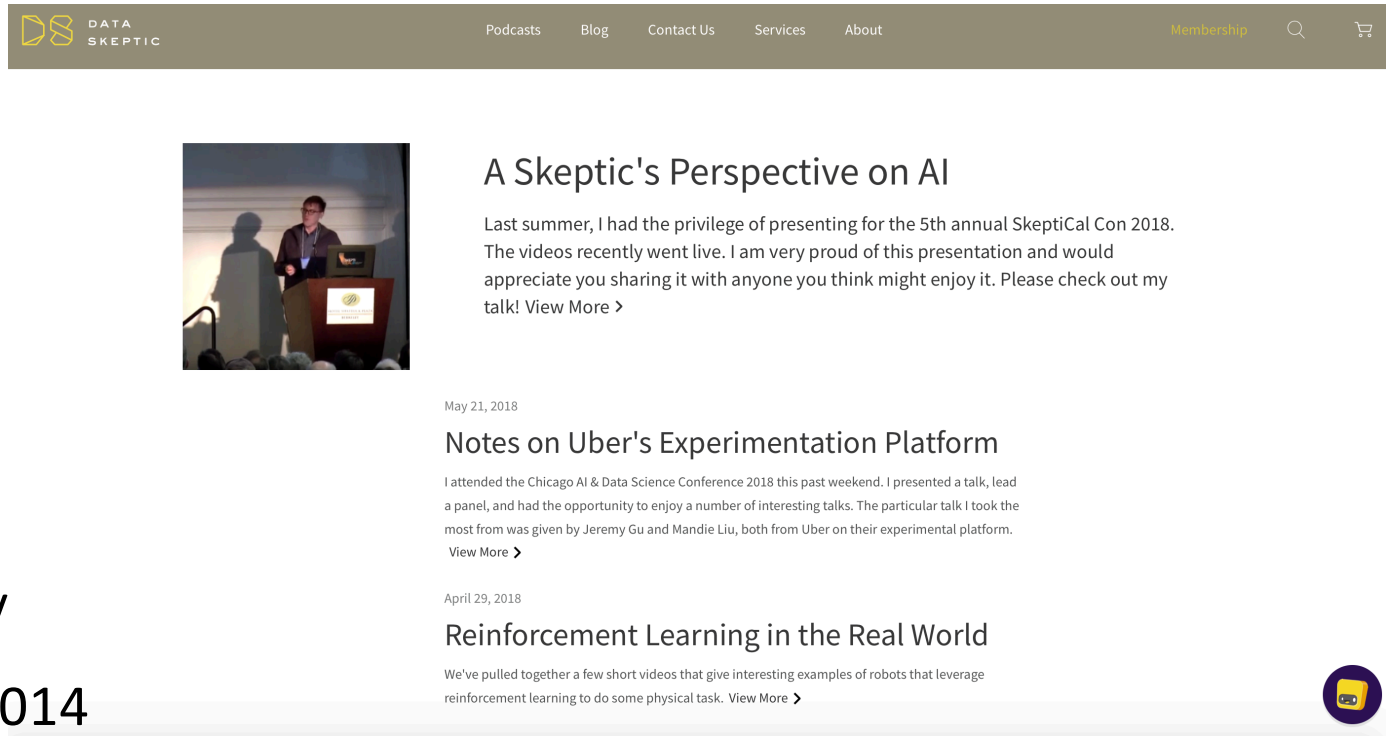
Open Access Books

- 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.





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)