

DSC 207 - PYTHON FOR DATA SCIENCE

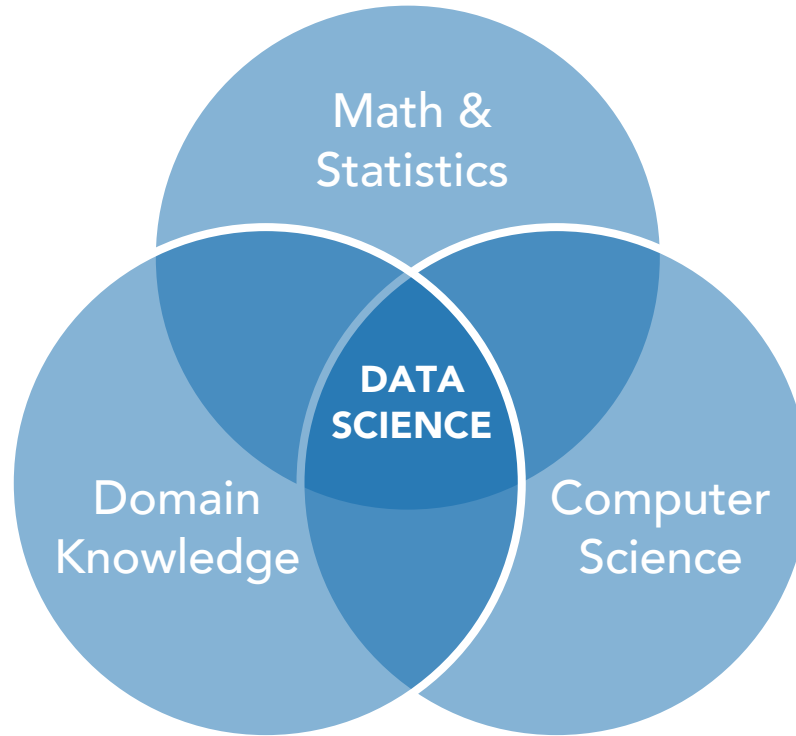
WHAT IS A DATA SCIENTIST?

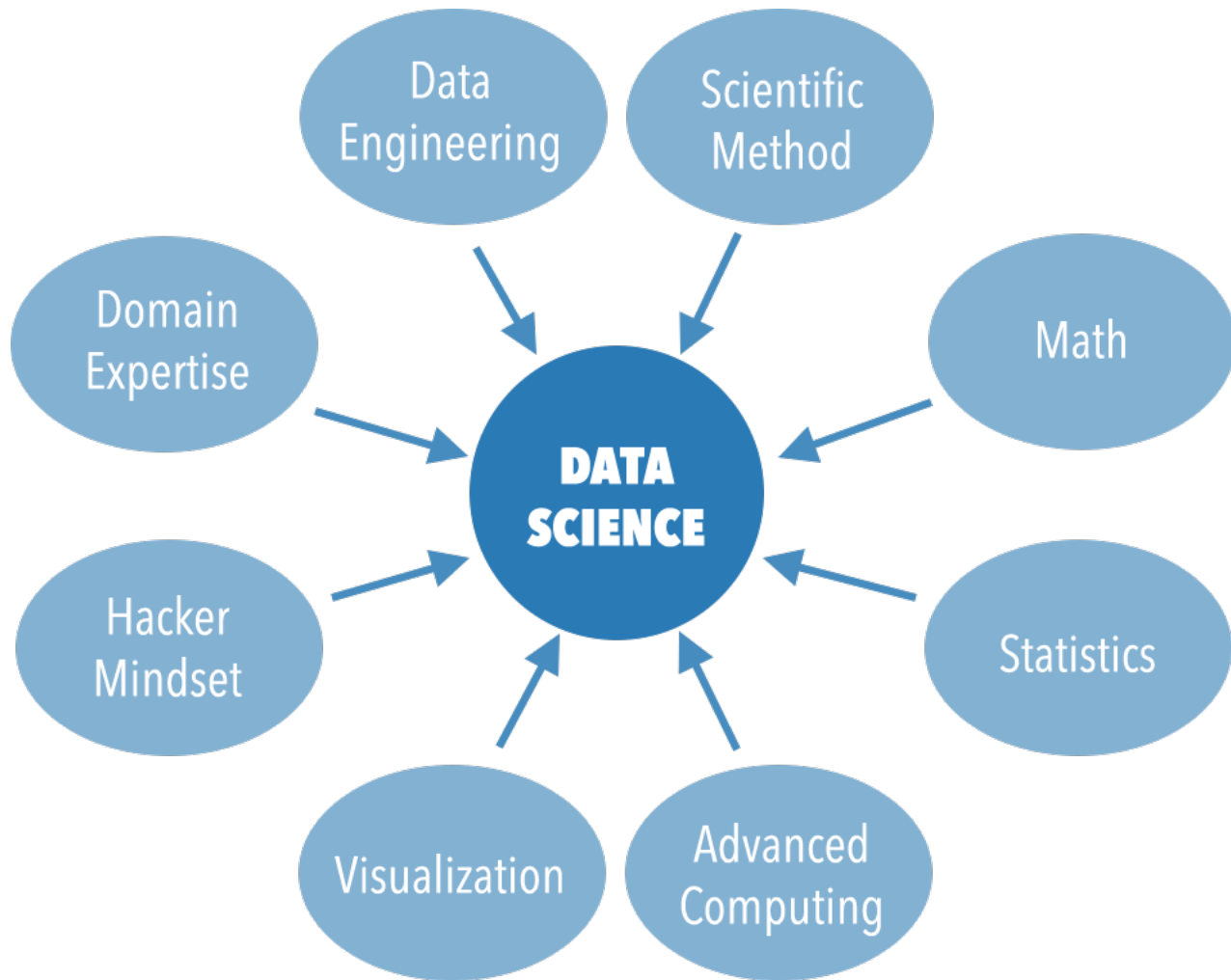
İLKAY ALTINTAŞ, PH.D.

By the end of this video, you will be able to:

- Explain the foundational skills for a data scientist
- Describe why Python is a required skill for data science today
- Gain exposure Jupyter notebooks and JupyterLab

Data science is multidisciplinary





MODERN DATA SCIENTIST

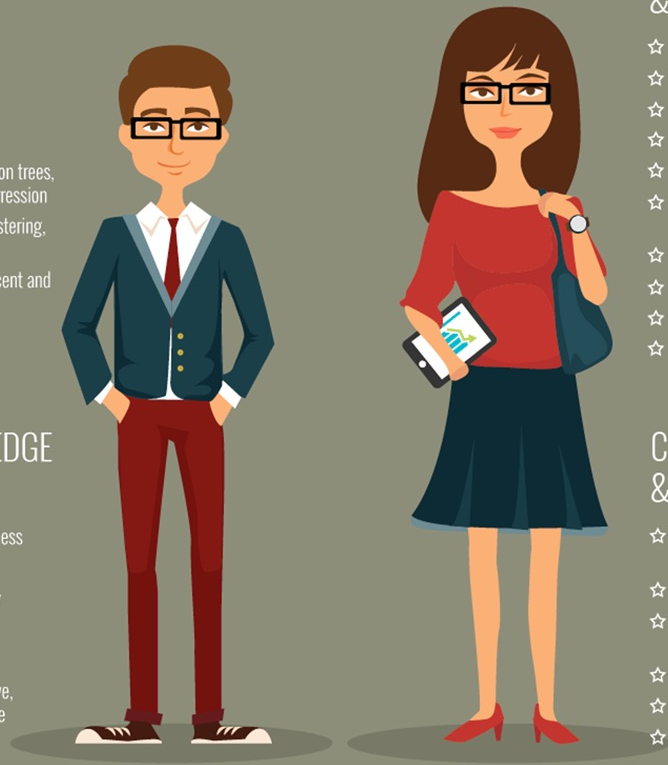
Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. 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 packages, 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

Are data scientists unicorns?



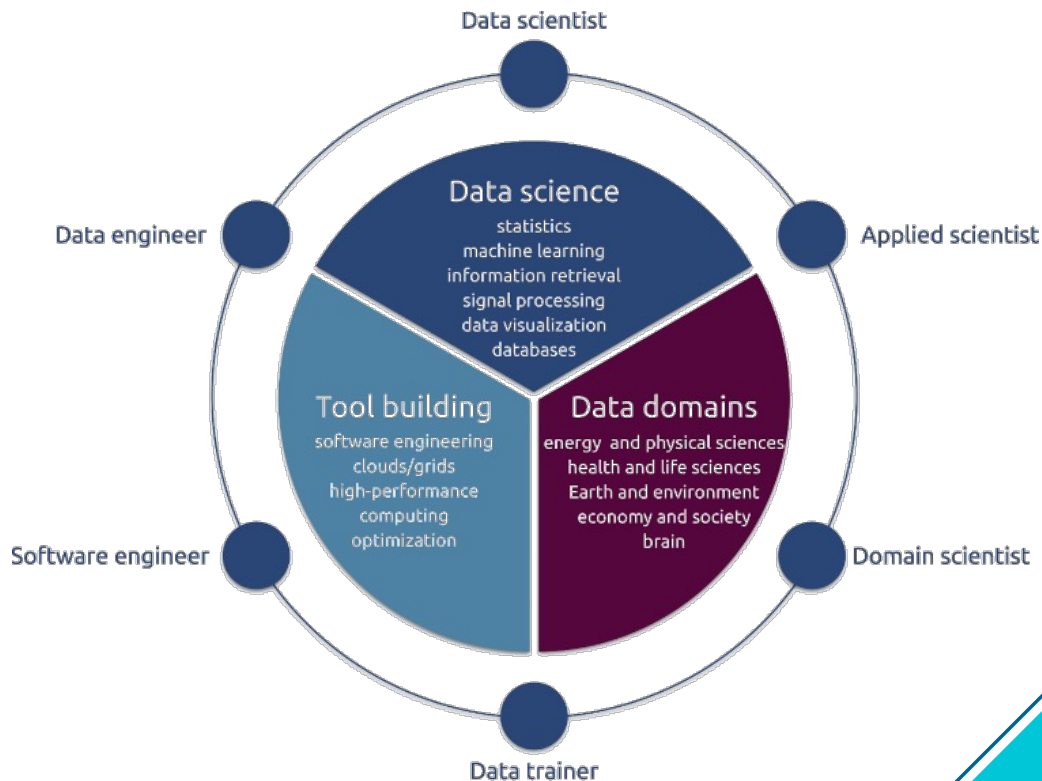
Data science is "WE" science!

- Team collaboration
- Curiosity
- Respect



Expertise and skills often overlap between different data science roles

- Data engineer
- Data analyst
- Methods expert
- Scalability and operations expert
- Business manager
- Business analyst
- Scientist
- Visualization and dashboard developer
- Solution architect
- Storyteller/coordinator
- Project manager



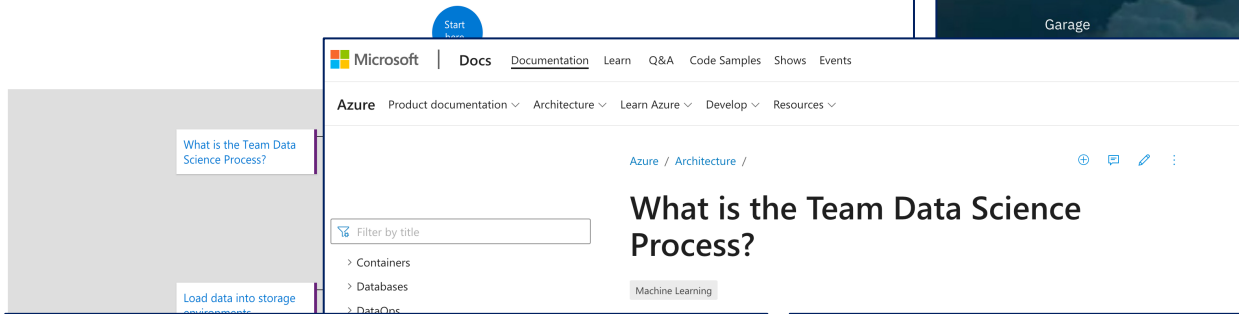
The data science ecosystem: activities and actors

<https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013>

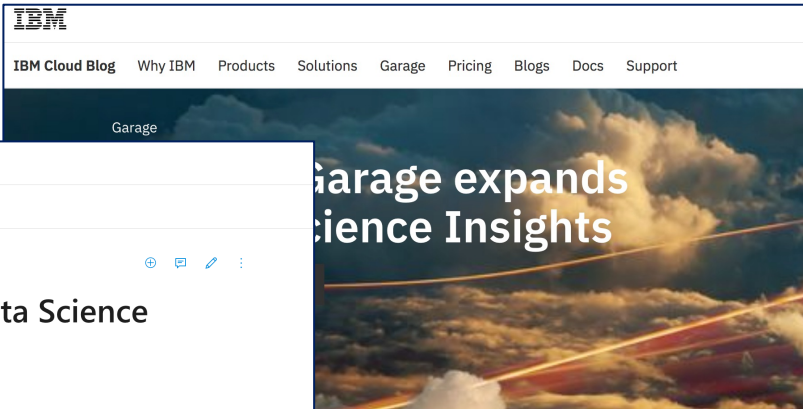
Team Data Science Environments

Using the Team Data Science Process with Azure Machine Learning

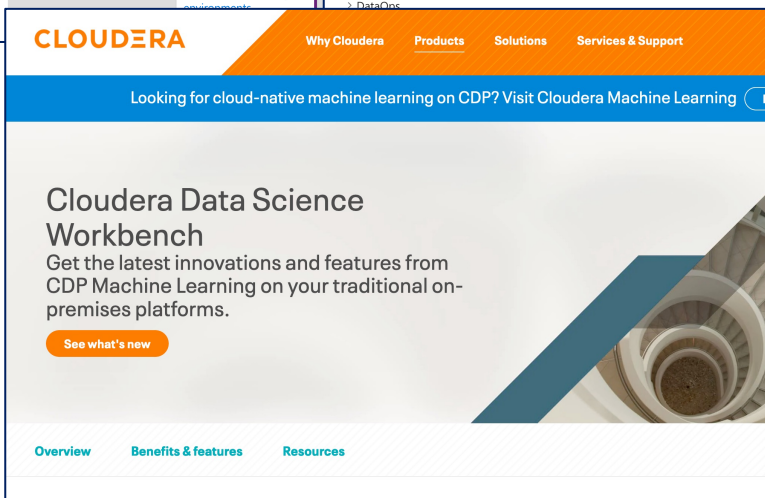
The Team Data Science Process (TDSP) provides a systematic approach to building intelligent applications that enables teams of data scientists to collaborate effectively over the full lifecycle of activities needed to turn these applications into products.



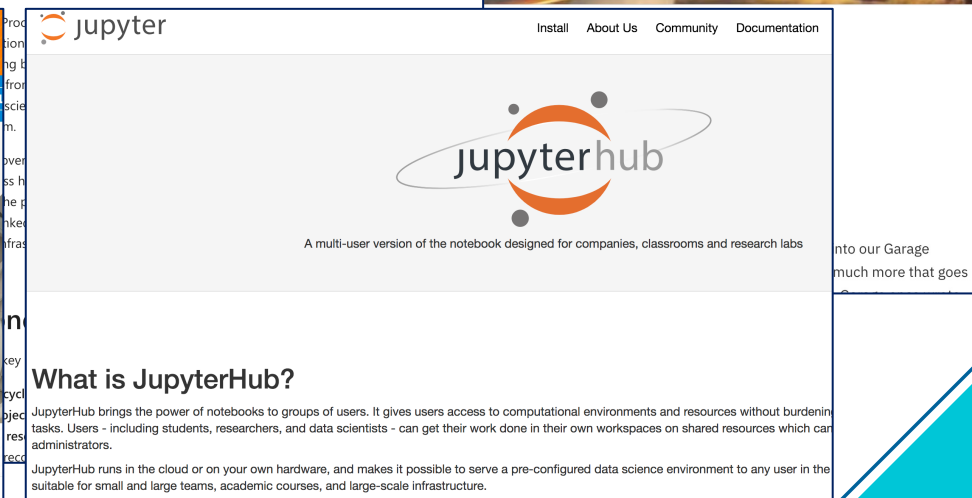
The screenshot shows the Microsoft Azure documentation website. The main heading is "What is the Team Data Science Process?". Below the heading, there is a "Filter by title" search bar and a list of categories: "Containers", "Databases", and "DataOps". A "Machine Learning" tag is visible. The page is part of the "Azure / Architecture /" section.



The screenshot shows the IBM Garage website. The main heading is "Garage expands data science insights". The page features a large image of a sunset over a body of water. The navigation bar includes links for "IBM Cloud Blog", "Why IBM", "Products", "Solutions", "Garage", "Pricing", "Blogs", "Docs", and "Support".



The screenshot shows the Cloudera Data Science Workbench website. The main heading is "Cloudera Data Science Workbench". Below the heading, there is a subheading "Get the latest innovations and features from CDP Machine Learning on your traditional on-premises platforms." and a button "See what's new". The page is part of the "Cloudera" brand.



The screenshot shows the JupyterHub website. The main heading is "What is JupyterHub?". Below the heading, there is a subheading "A multi-user version of the notebook designed for companies, classrooms and research labs". The page features the JupyterHub logo and a description of the platform. The navigation bar includes links for "Install", "About Us", "Community", and "Documentation".

Top Data Science Programming Languages

Platform	2019 % share	2018 % share	% change
Python	65.8%	65.6%	0.2%
R Language	46.6%	48.5%	-4.0%
SQL Language	32.8%	39.6%	-17.2%
Java	12.4%	15.1%	-17.7%
Unix shell/awk	7.9%	9.2%	-13.4%
C/C++	7.1%	6.8%	3.7%
Other programming and data languages	6.8%	6.9%	-17.1%
Scala	3.5%	5.9%	-41.0%
Julia	1.7%	0.7%	150.4%
Perl	1.3%	1.0%	25.2%
Lisp	0.4%	0.3%	46.1%
Javascript	6.8%	na	na

Why Python?



Getting Started

Python can be easy to pick up whether you're a first time programmer or you're experienced with other languages. The following pages are a useful first step to get on your way writing programs with Python!

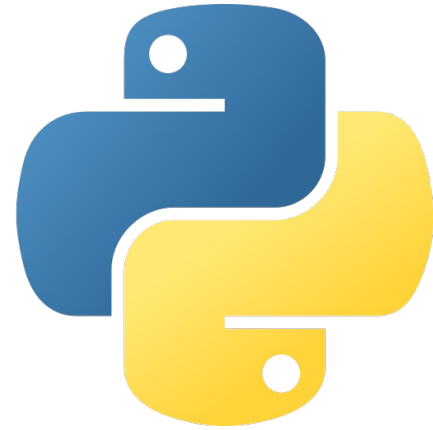
Friendly & Easy to Learn

The community hosts conferences and meetups, collaborates on code, and much more. Python's documentation will help you along the way, and the mailing lists will keep you in touch.

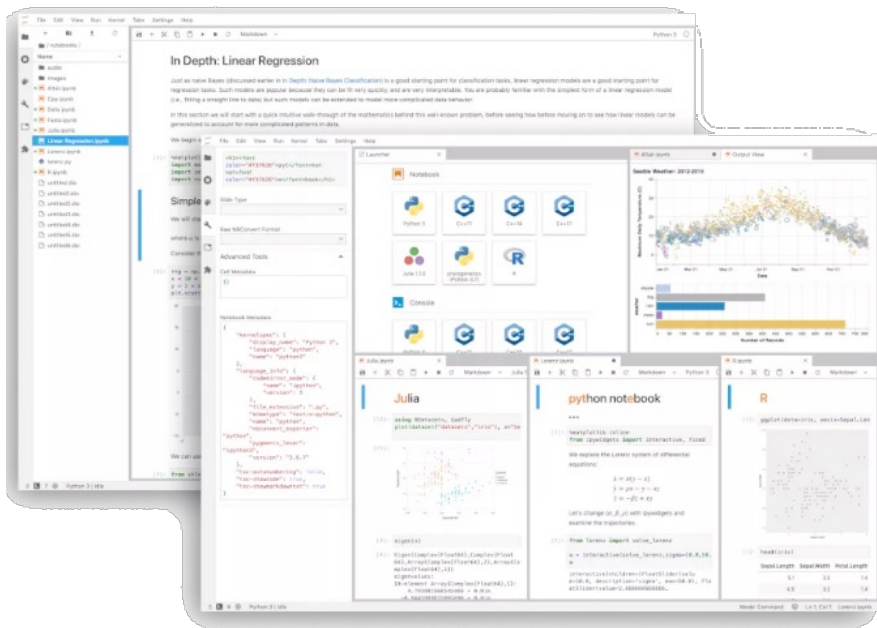
<https://www.python.org/about/>

Why Python for Data Science?

- Easy-to-read and learn
- Vibrant community
- Growing and evolving set of libraries
 - Data management
 - Analytical processing
 - Visualization
- Applicable to each step in the data science process
- Jupyter Notebooks



Jupyter Notebooks and JupyterLab



JupyterLab: A Next-Generation Notebook Interface

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

Try it in your browser

Install JupyterLab



<https://jupyter.org/>