

Data Science and the Data Scientist Toolkit



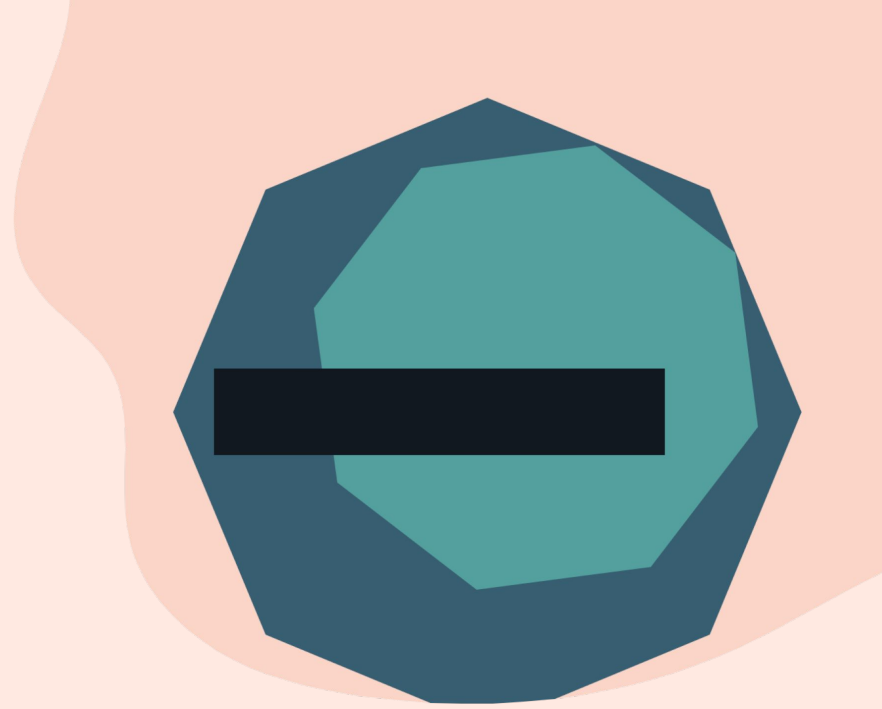
// FLATIRON SCHOOL

Agenda

A large, teal-colored polygon is positioned on the left side of the slide, partially overlapping the dark blue background. It has several vertices, creating a complex, angular shape.

- What is Data Science?
 - Roles and Responsibilities
 - The Process
- The Data Science Toolkit (Phase 1)

**So:
What is
Data Science?**



What is Data Science?

Find out for yourself!



Prompt: Spend the next 10 minutes skimming and discussing your assigned blog post, then come back and report your findings to the rest of us.

1. [A Deep Look Into 13 Data Scientist Roles and Their Responsibilities](#)
2. [The Data Science Process](#)
3. [Most In Demand Data Science Technical Skills](#)
4. [A Learning Path to Becoming a Data Scientist](#)
5. [Compilation of Advice for New and Aspiring Data Scientists](#)

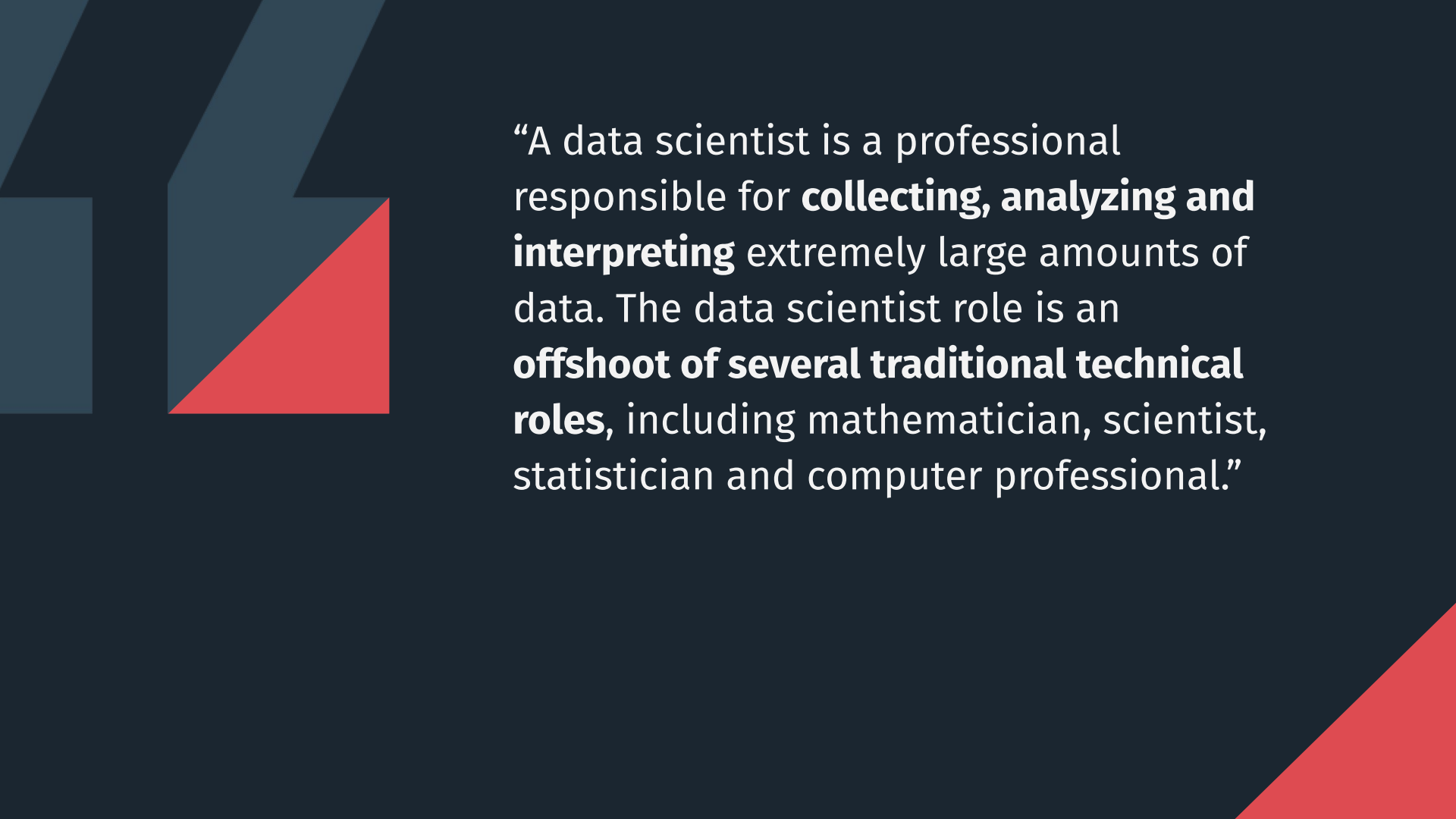


Let's Discuss!

What does a “data scientist” do?

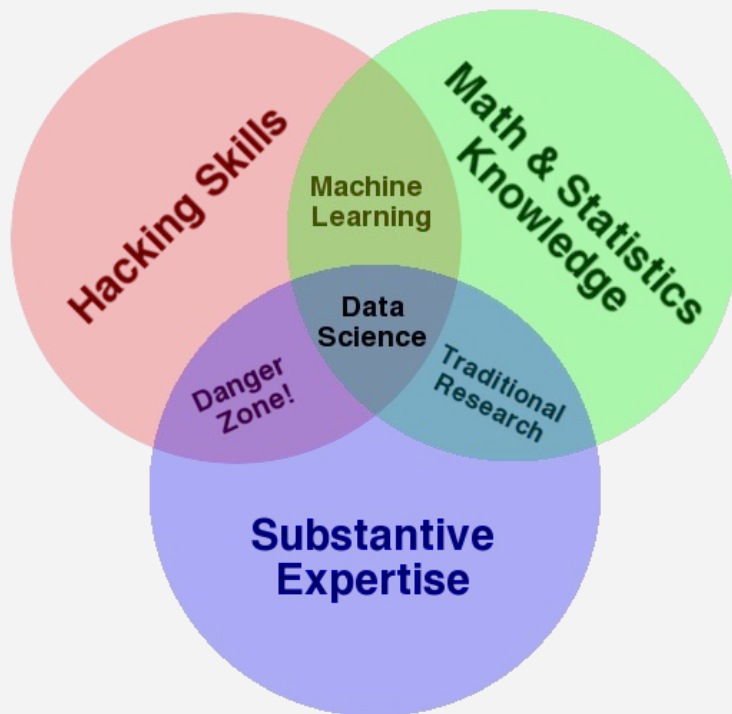
What are the main skills you need to be a “data scientist” ?

What is consistent among these posts, and what is in dispute?

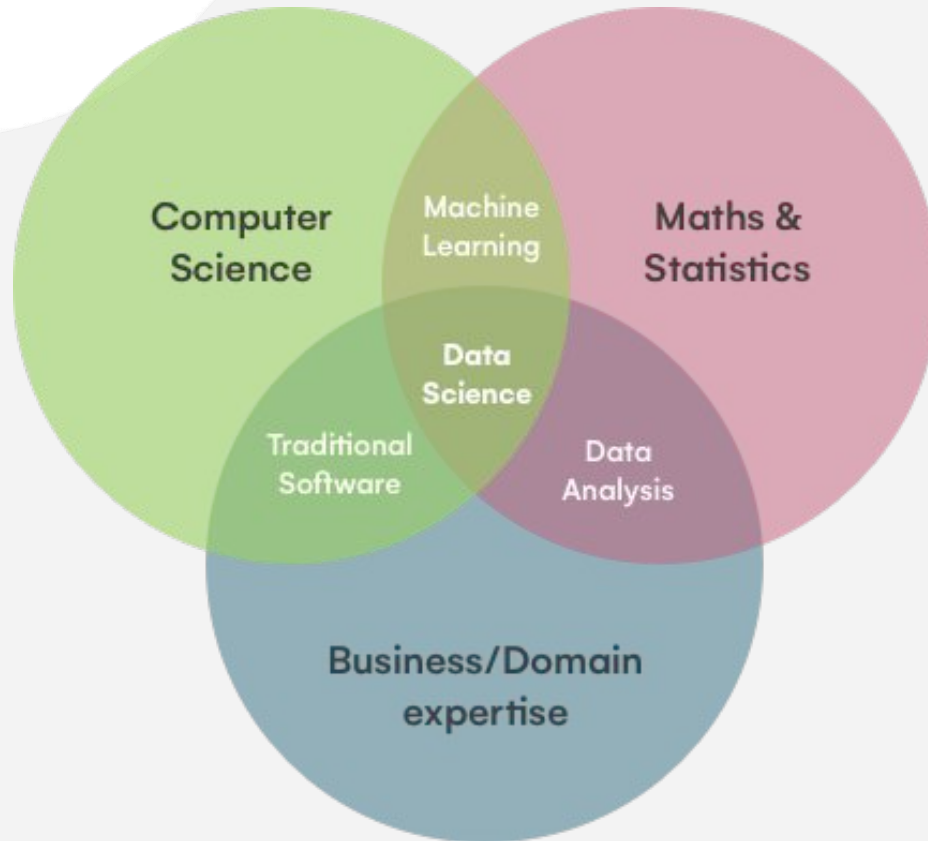
The background features a dark navy blue field. On the left side, there are several overlapping geometric shapes: a dark blue square, a dark blue rectangle, and a red triangle pointing towards the center. A larger red triangle is also visible in the bottom right corner of the frame.

“A data scientist is a professional responsible for **collecting, analyzing and interpreting** extremely large amounts of data. The data scientist role is an **offshoot of several traditional technical roles**, including mathematician, scientist, statistician and computer professional.”

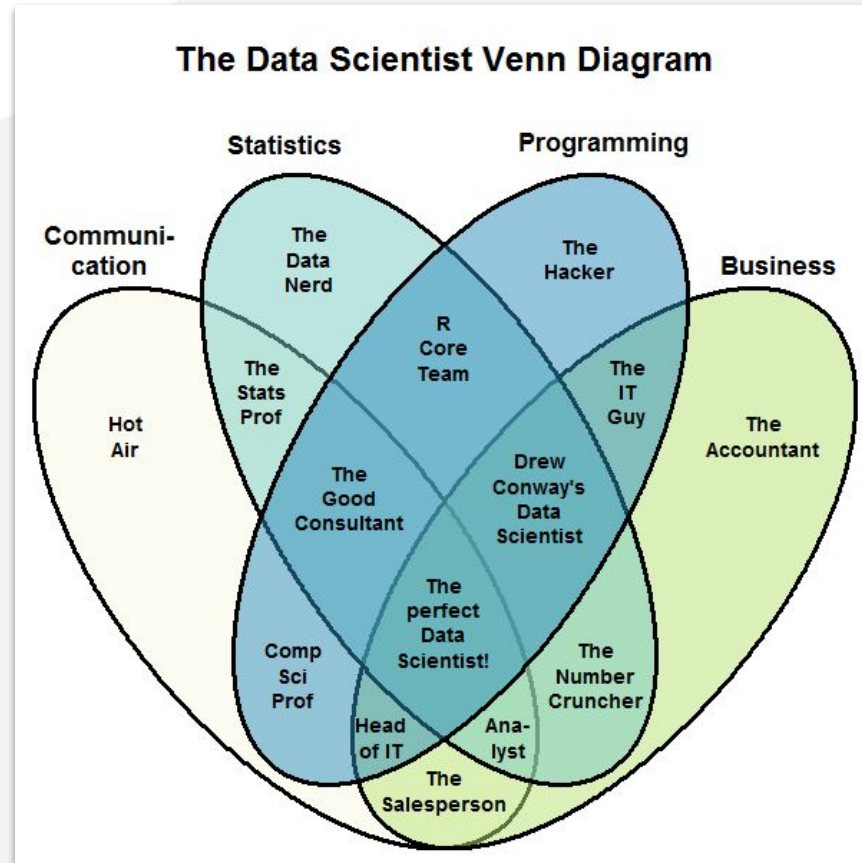
The Data Science Venn Diagram



Another Version




And Another




Common Roles & Responsibilities

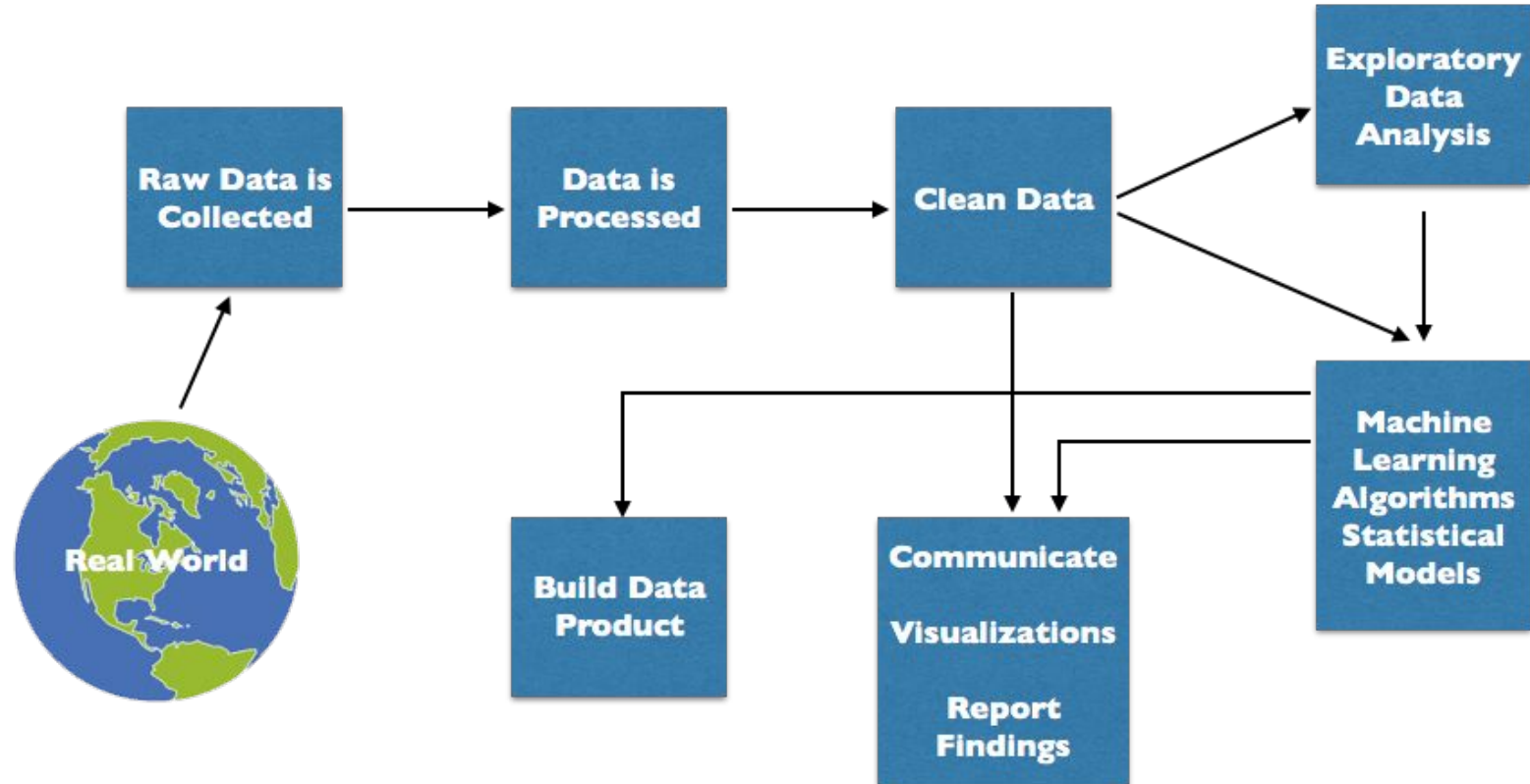
	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	Very important	Very important	Very important	Very important
Data Visualization and Communication	Very important	Somewhat important	Somewhat important	Very important
Data Intuition	Somewhat important	Very important	Somewhat important	Very important
Statistics	Somewhat important	Very important	Somewhat important	Very important
Data Wrangling	Not that important	Not that important	Very important	Very important
Machine Learning	Not that important	Very important	Not that important	Very important
Software Engineering	Not that important	Somewhat important	Very important	Somewhat important
Multivariable Calculus and Linear Algebra	Not that important	Very important	Not that important	Somewhat important
<div><div></div> Not that important</div> <div><div></div> Somewhat important</div> <div><div></div> Very important</div>				



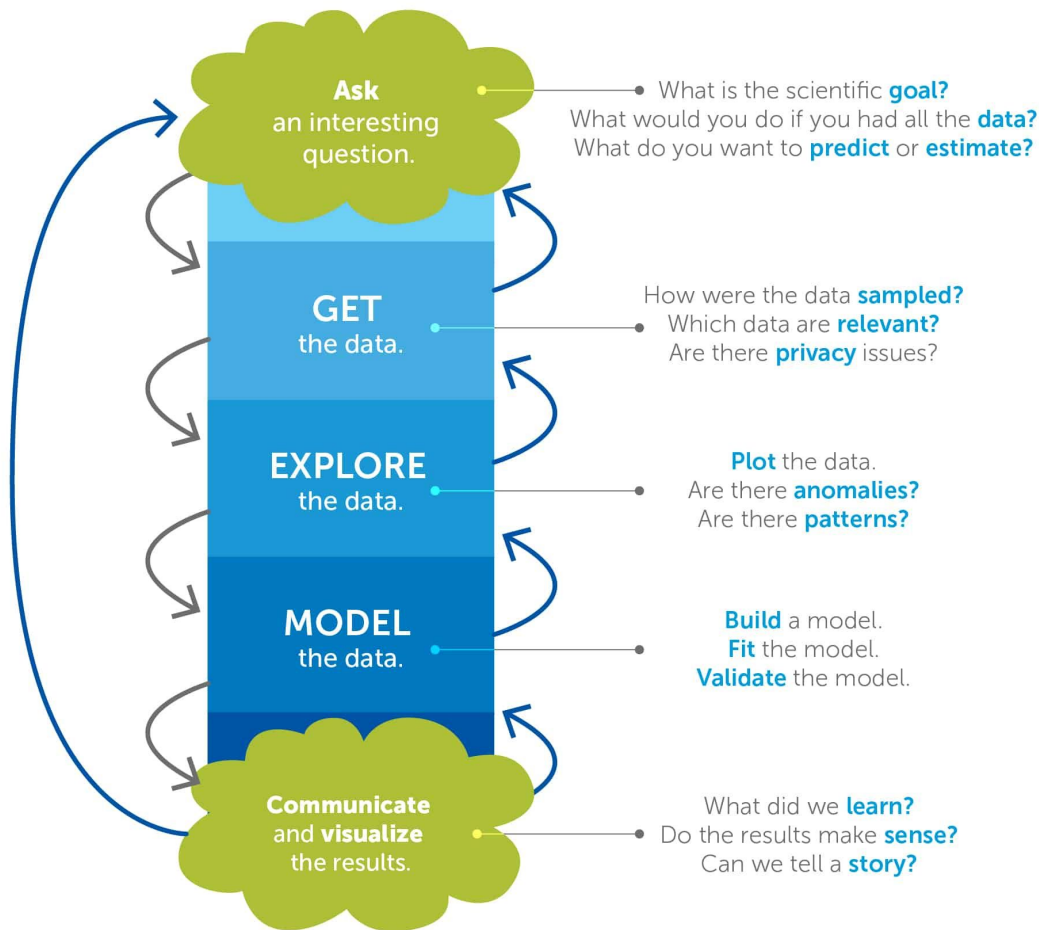
“Regardless of your exact job title, if you’re in the field of data science, you’ll be expected to be **involved in a lot of different steps** in the data-driven product development cycle. You should be ready to discover new areas to **optimize**, figure out the **metrics** that matter, find the **data** to inform these metrics, design and execute **experiments**, and **present the results** of experiments/models in concise, accurate, and convincing ways.”



The Data Science Process

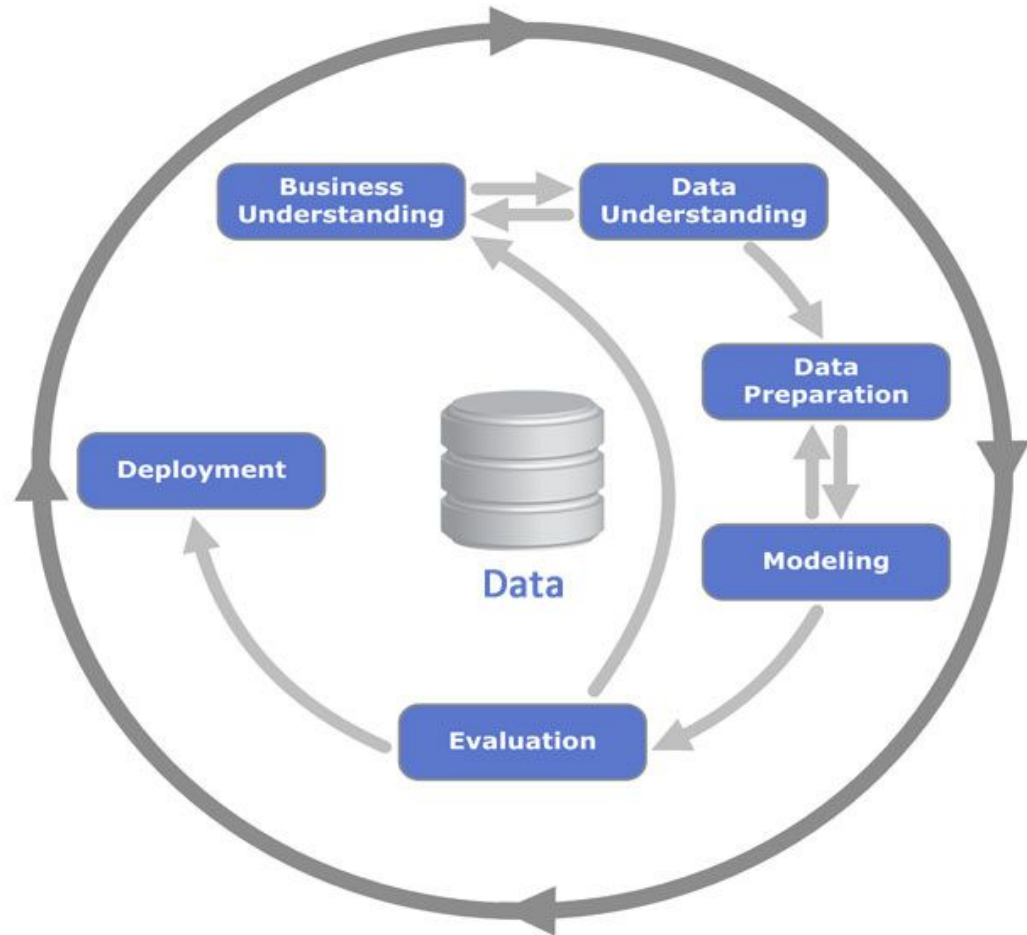


And Another



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.

CRISP-DM Process Diagram



Source: Kenneth Jensen

Data Science Process



OBTAIN



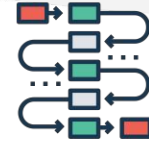
SCRUB



EXPLORE



MODEL



INTERPRET

O

Gather data from relevant sources

S

Clean data to formats that machine understands

E

Find significant patterns and trends using statistical methods

M

Construct models to predict and forecast

N

Put the results into good use

A dark blue, solid-colored octagon is positioned in the upper right quadrant of the slide. It is partially overlaid by a diagonal line that separates a light peach background from a darker teal background in the top right corner.

The Data Science Toolkit

Data Science Toolkit - Phase 1

Languages



Interfaces



Version Control



Package Control



Languages



Python

- Free, open source, versatile, powerful
- Not just for data science!
- Object-oriented (everything is an 'object')
- [The Zen of Python](#)



Structured Query Language (SQL)

- Connect to, change, and retrieve data from relational databases
- Developed in the 1970s, still going strong
- Many flavors

Interfaces



Jupyter Notebooks

- Streamlined document-centric interface for running and sharing code



IllumiDesk

- Hosts Jupyter Notebooks in the cloud



Code-Focused Text Editor

- Write text files in a code-native format
- **VS Code** is one of many that would work

Version Control



Git

- Distributed version tracking on any files
- Folder → “Repository”



GitHub

- Hosts Git repositories
- Collaborate and share code with others
- Backbone of the open source community
- Your Data Science portfolio!

Package Control



Anaconda

- Package management and deployment
- Designed with Data Science in mind
- Create and share environments



Python Package Index (PyPi)

- Database of public Python libraries
- Package installer (pip)
- Not everything is on Anaconda

The background is a dark navy blue. In the top-left corner, there are several overlapping geometric shapes in a medium blue-grey color. One of these shapes is a square, and another is a larger shape that includes a red triangle at its bottom-right corner. This red triangle is a right-angled isosceles triangle. In the bottom-right corner, there is a large red triangle pointing towards the center of the slide.

**Now:
Time to Get
Started!**