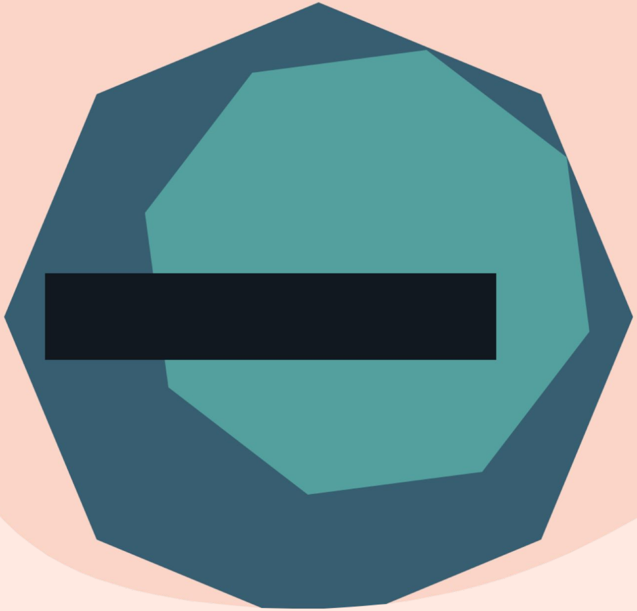
Data Science and the Data Scientist Toolkit

// FLATIRON SCHOOL

Agenda

- 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!

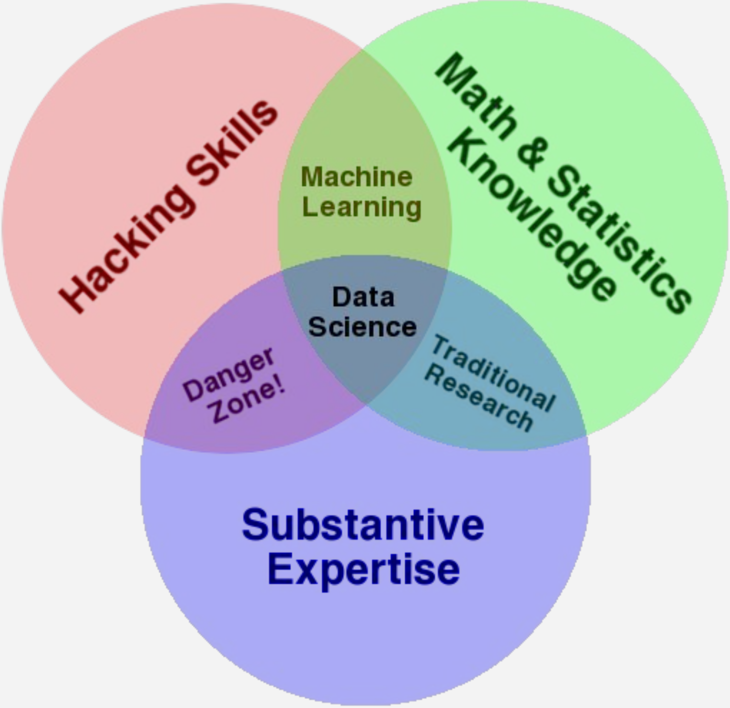
What does a “data scientist” do?

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

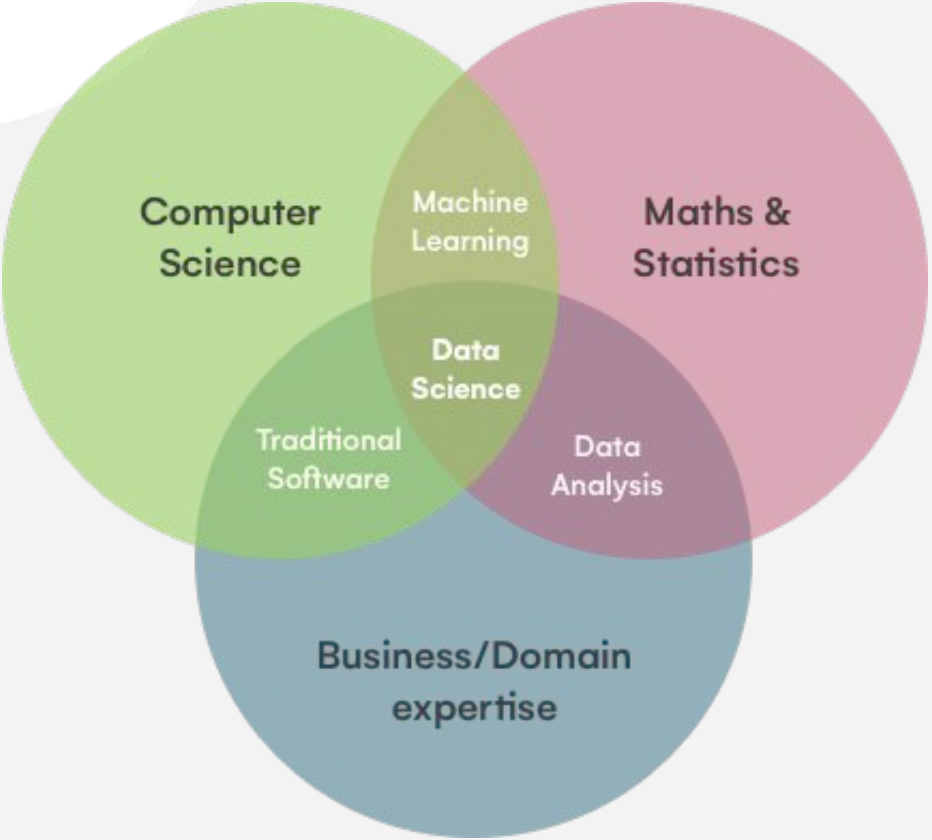
What is consistent among these posts, and what is in dispute?

“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.”

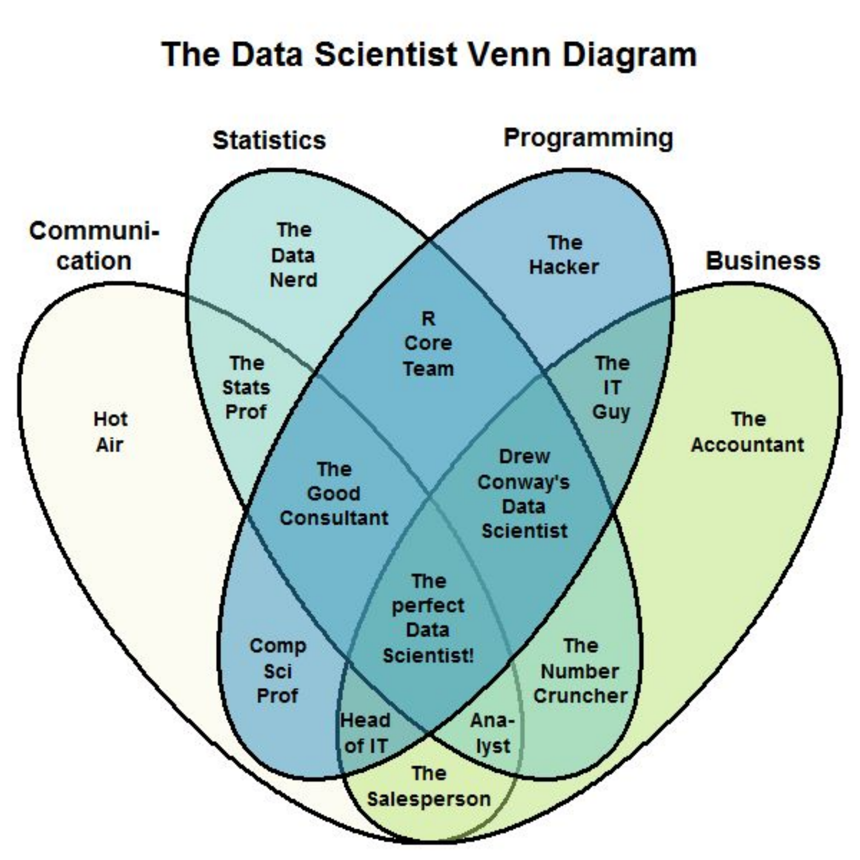
The Data Science Venn Diagram



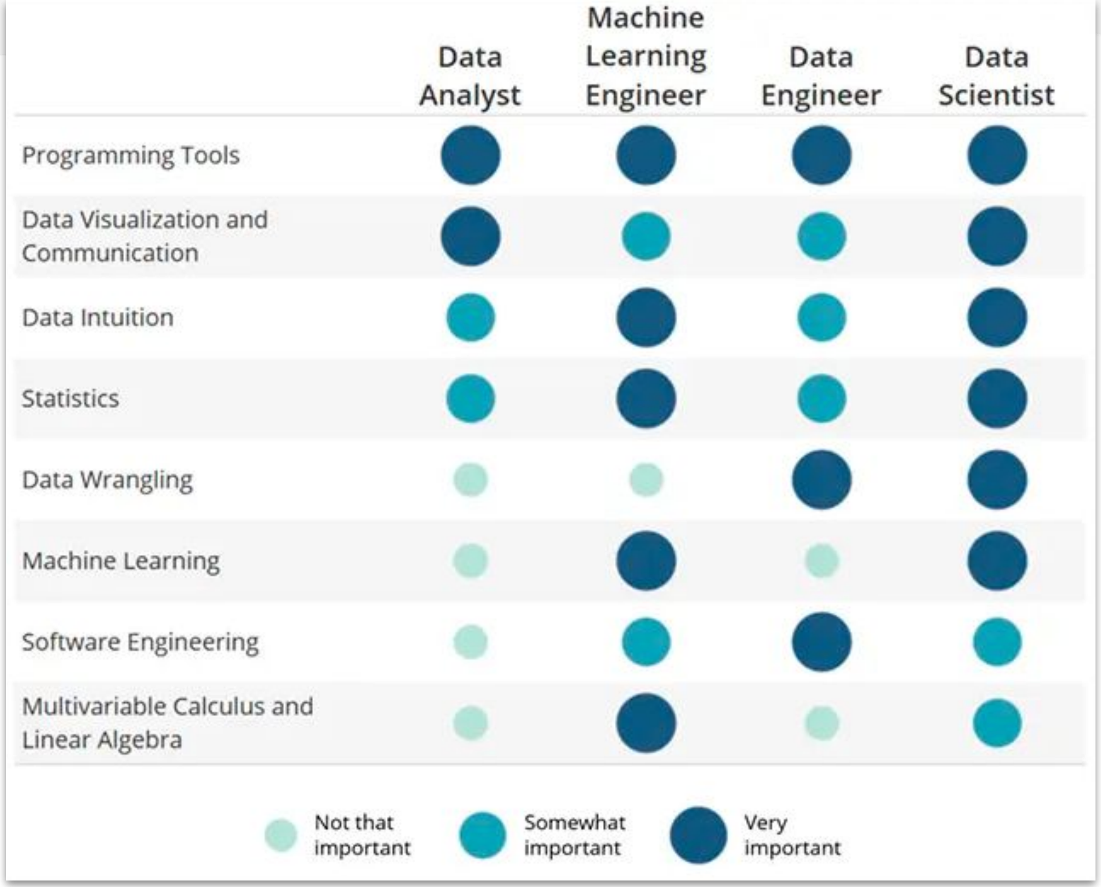
Another Version



And Another


The Data Scientist Venn Diagram
Statistics
Programming
Communi-cation
Business
The Data Nerd
The Hacker
R Core Team
The Stats Prof
The IT Guy
Hot Air
The Accountant
The Good Consultant
Drew Conway's Data Scientist
The perfect Data Scientist!
Comp Sci Prof
The Number Cruncher
Head of IT
Ana-lyst
The Salesperson

Common Roles & Responsibilities

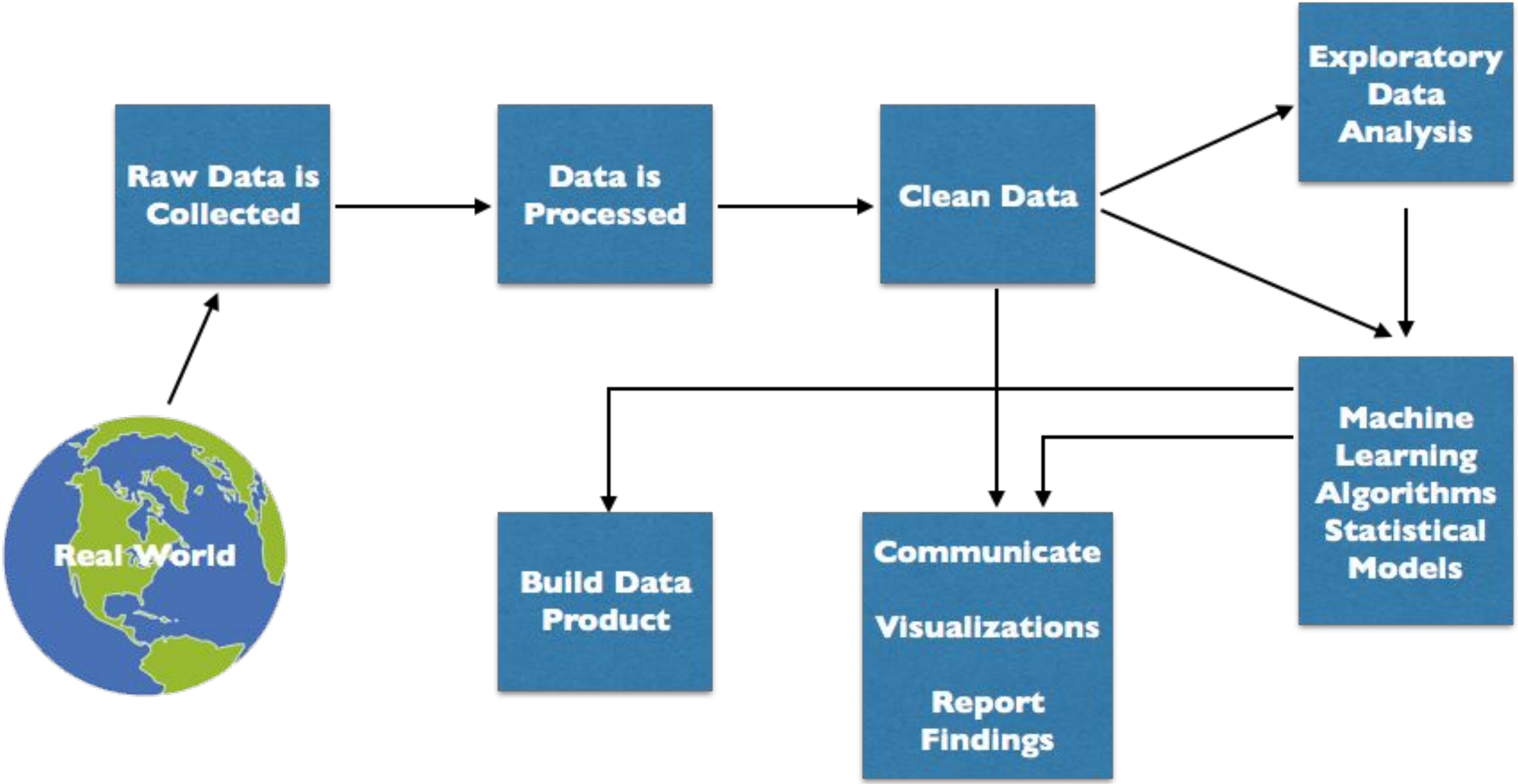


	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◐	◐	●
Data Intuition	◐	●	◐	●
Statistics	◐	●	◐	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◐	●	◐
Multivariable Calculus and Linear Algebra	○	●	○	◐

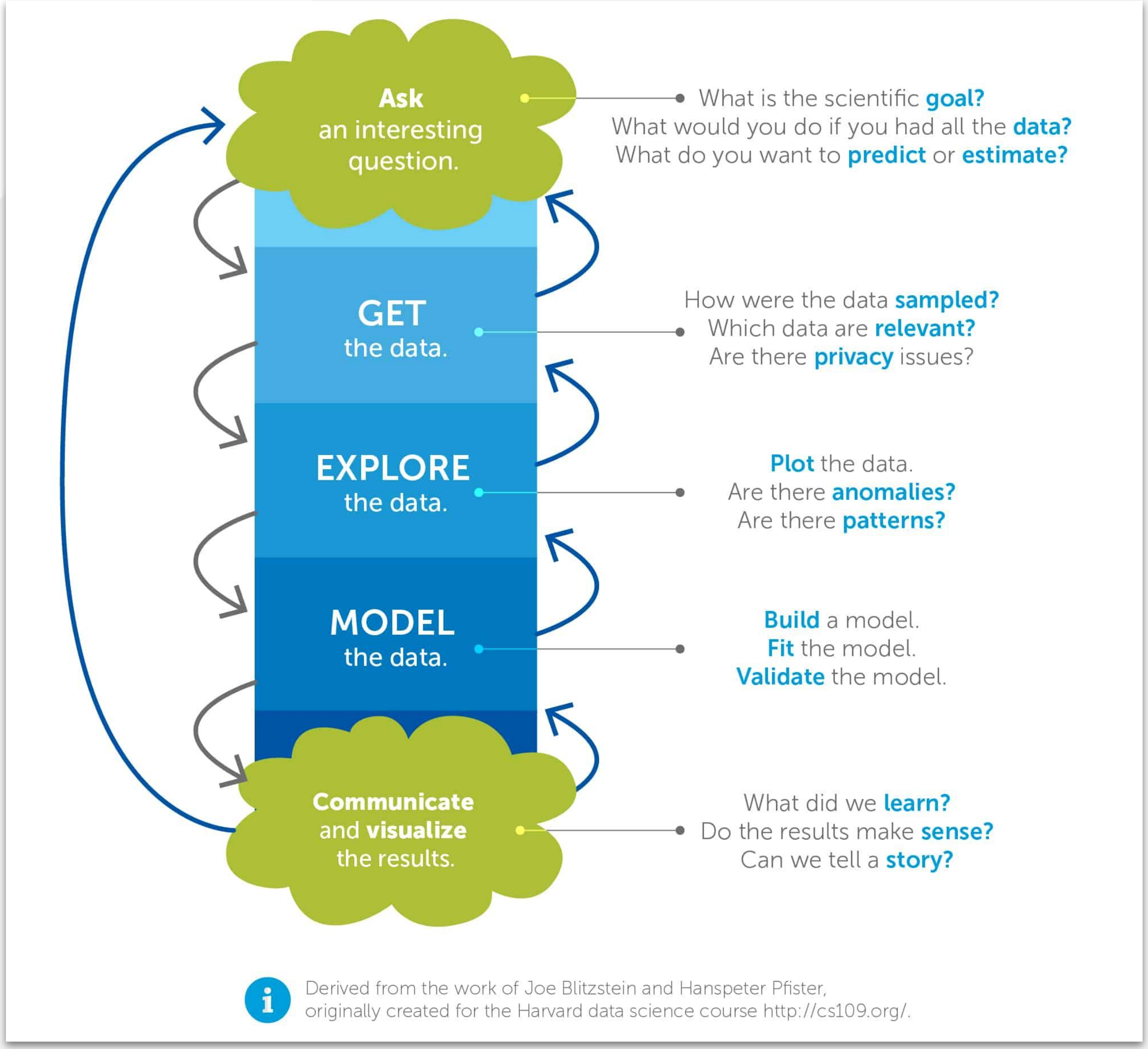
○ Not that important ◐ Somewhat important ● Very important

“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



Ask an interesting question.

What is the scientific **goal?**
What would you do if you had all the **data?**
What do you want to **predict** or **estimate?**

GET the data.

How were the data **sampled?**
Which data are **relevant?**
Are there **privacy** issues?

EXPLORE the data.

Plot the data.
Are there **anomalies?**
Are there **patterns?**

MODEL the data.

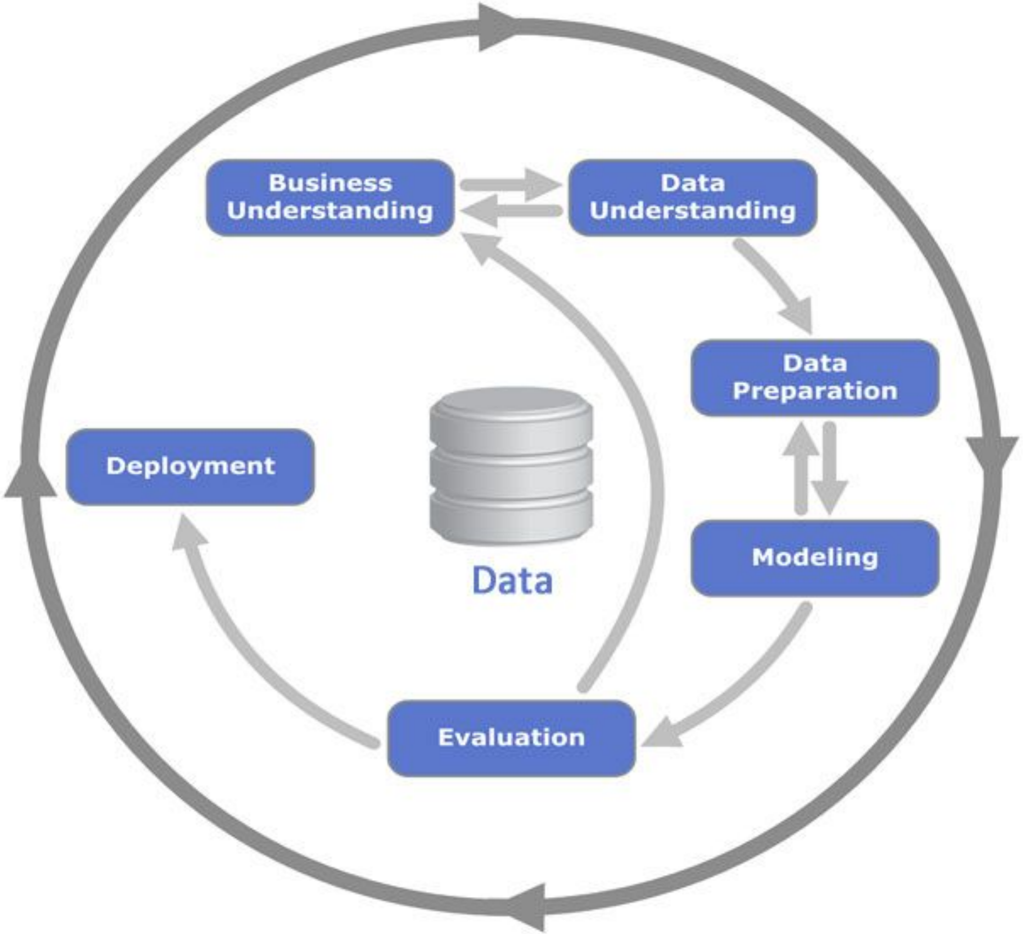
Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

What did we **learn?**
Do the results make **sense?**
Can we tell a **story?**

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



Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

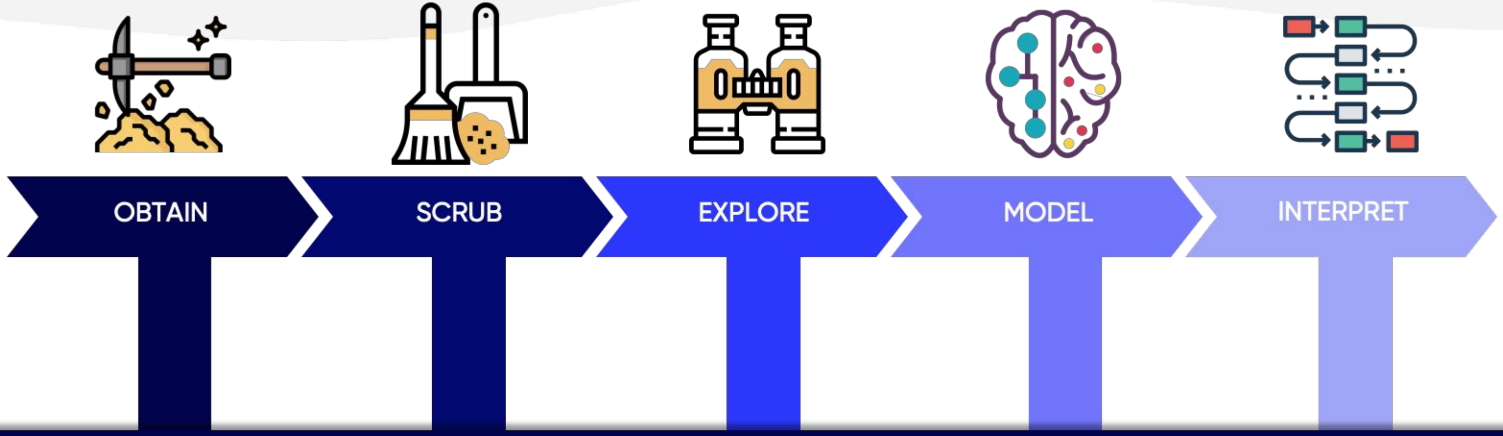
Deployment

Data

Source: Kenneth Jensen

Data Science Process

LEAD



OBTAIN

O

Gather data from relevant sources

SCRUB

S

Clean data to formats that machine understands

EXPLORE

E

Find significant patterns and trends using statistical methods

MODEL

M

Construct models to predict and forecast

INTERPRET

N

Put the results into good use

Originally by Hilary Mason and Chris Wiggins

The Data Science Toolkit

Data Science Toolkit - Phase 1

Languages

python™

SQL

Interfaces

jupyter

SaturnCloud

Version Control

git

GitHub

Package Control

ANACONDA.

pypi

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

Now:
Time to put this all to use!