

What Is Data Science?

(aka what is this buzzword anyway?)





Today's Learning Objectives

- Explain some of the types of problems that data science can be used to solve
- Describe the data science process
- List a few types of skills that all conglomerate into 'data science'



Find out for yourself:

Pick one, read through it, summarize it, share with the class:

1. [Battle of the Data Science Venn Diagrams](#) (blog post, 2 pages)
2. [The 5 Questions Data Science Can Answer](#) (5 min video)
3. [Preparing for the Transition to Data Science](#) (blog post)
4. [How to Become a \(Good\) Data Scientist](#) (blog post)
5. [The Data Science Process](#) (blog post)



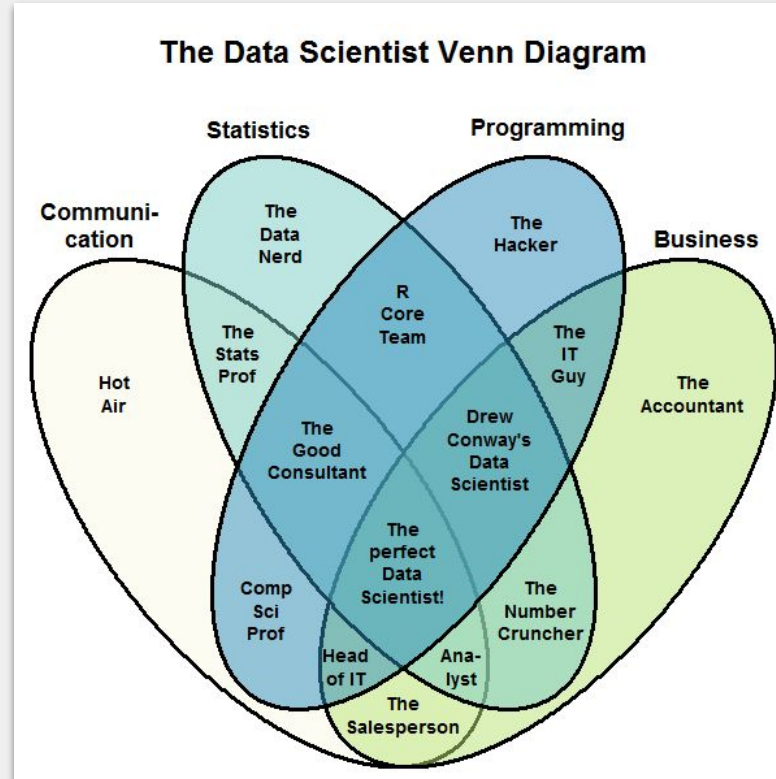
Let's discuss!

Across these articles/blogs/videos, what are the main skills you need to be a data scientist?

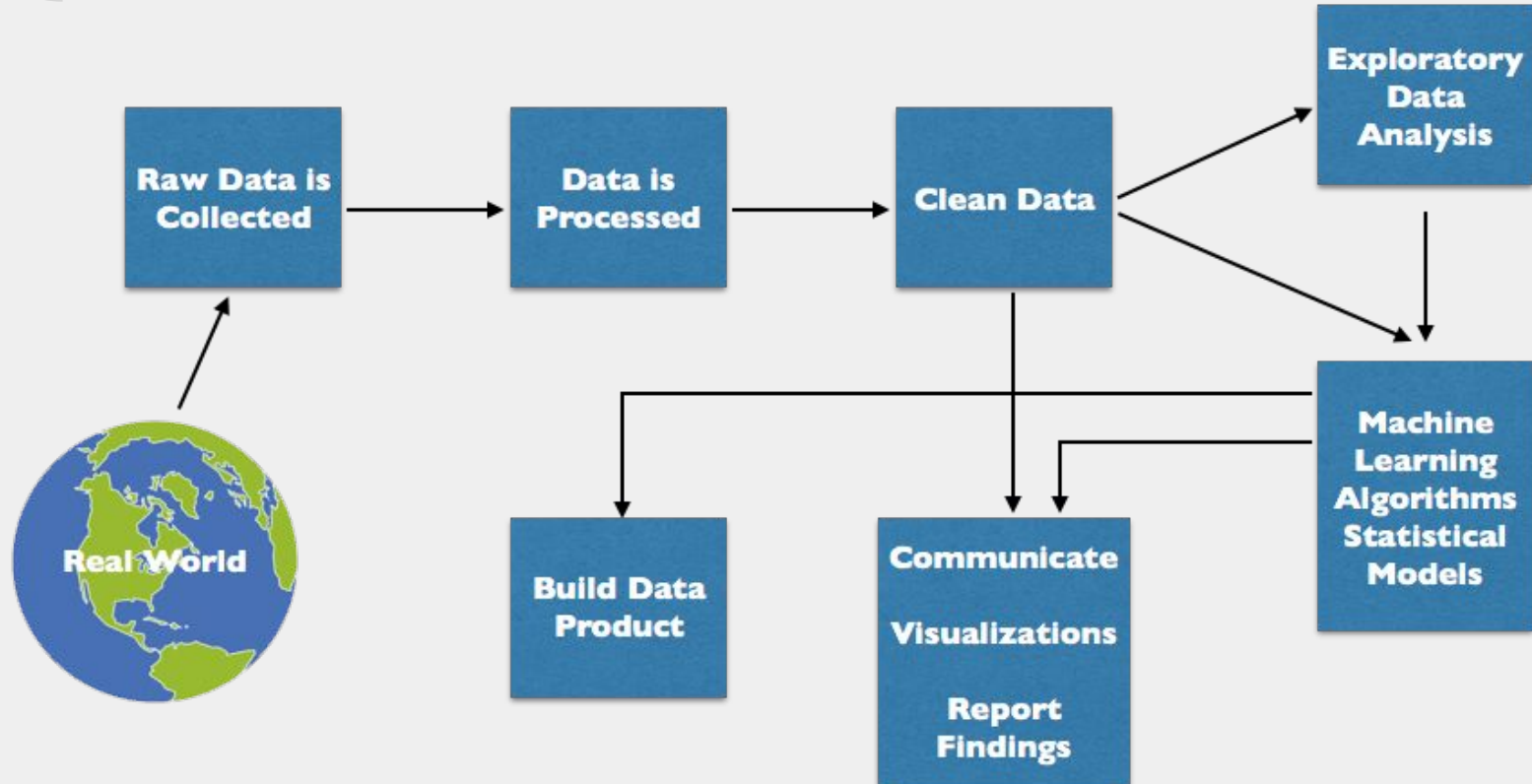
What things are in dispute? In other words, what isn't consistent across these sources, or what do some talk about that others don't?

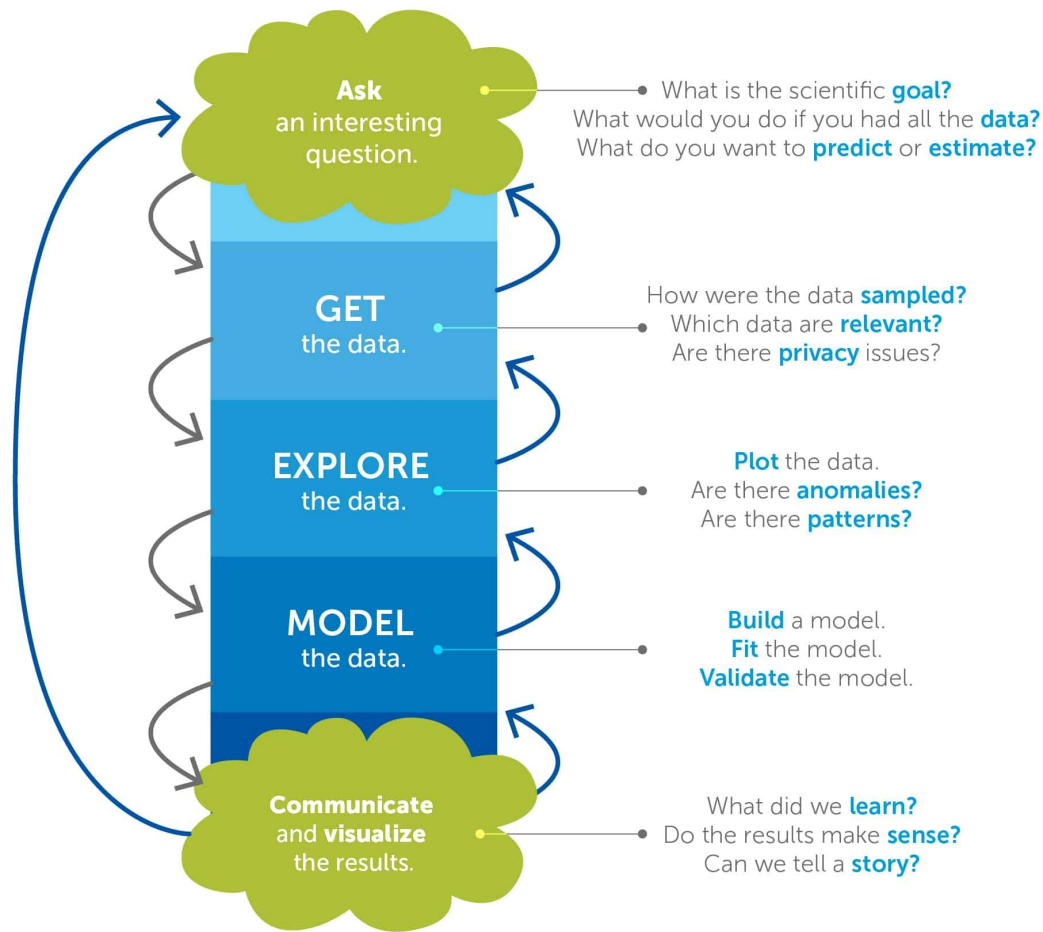
What does a 'data scientist' even do?

Overview



Data Science Process



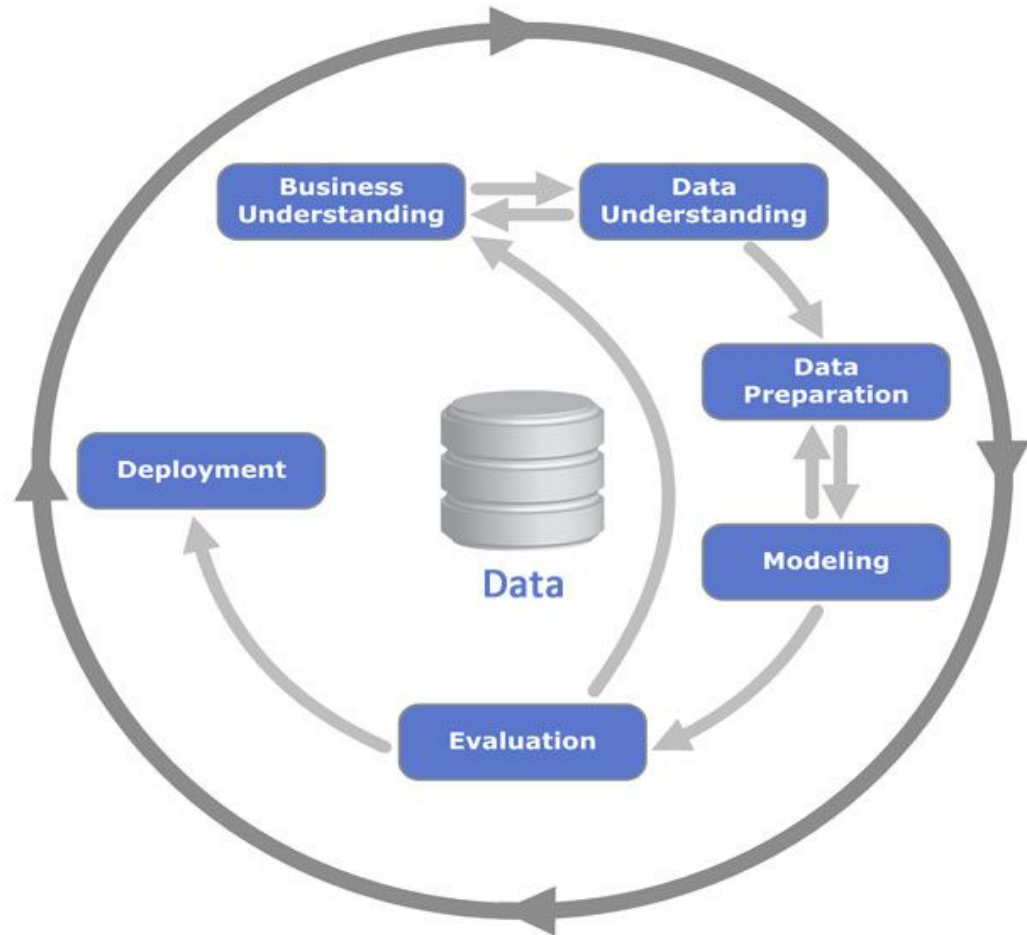


Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course <http://cs109.org/>.

Data Science Deconstructed



CRISP-DM Process Diagram



Source: Kenneth Jensen

Data Science Process



OBTAIN



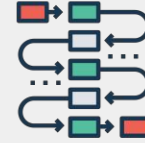
SCRUB



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



Takeaways

- Question First versus Data First
- Data Cleaning/Preparation/Processing/Exploration is key
- Communication (of question and of results) is also key
- Have to ask questions that models/machines can answer
- The process is iterative: you don't just finish a step, but go back and forth repeatedly