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. <u>Battle of the Data Science Venn Diagrams</u> (blog post, 2 pages)
- 2. The 5 Questions Data Science Can Answer (5 min video)
- 3. <u>Preparing for the Transition to Data Science</u> (blog post)
- 4. How to Become a (Good) Data Scientist (blog post)
- 5. <u>The Data Science Process</u> (blog post)

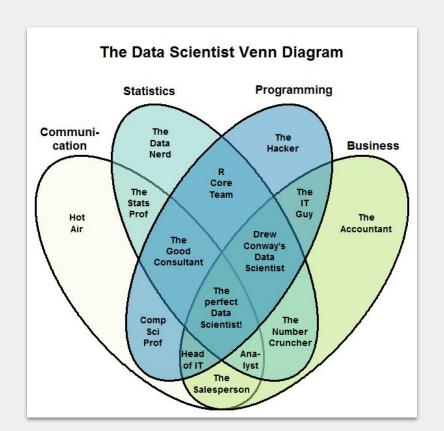
Let's discuss!

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

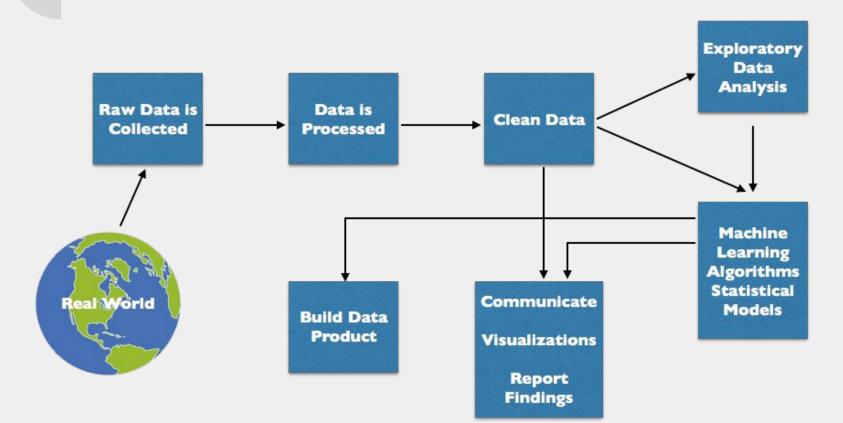
What things are in dispute? In other words, what isn't consistent across these sources, or what do some talk about that others don't?

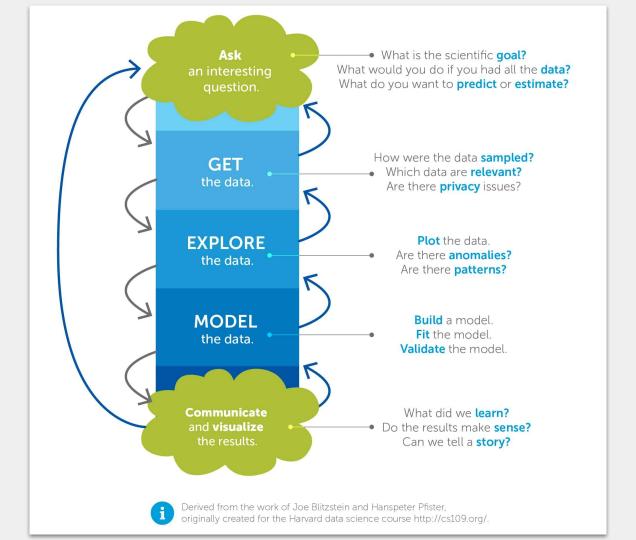
What does a 'data scientist' even do?

Overview



Data Science Process





Data Science Deconstructed

Ask a Lot of Questions

- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

01 D 02 Frame the Collect

Identify All Available Datasets

Web, internal/ external databases, etc.

Extract Data Into Usable Format

.csv, .json, .xml, etc.

Identify Business Insights

Return back to the business problem

Visualize Your Findings

Keep it simple & priority-driven

Tell a Clear & Actionable Story

 Effectively communicate to non-technical audiences

06 Communicate Results

THE DATA SCIENCE PROCESS

03 Process the Data

05 Perform In-Depth Analysis

04 Explore the Data

Raw Data

Examine Data at a High-Level

 Understand every column; identify errors, missing values & corrupt records

Clean the data

 Throw away, replace, and/or filter corrupt /error prone / missing values

Create a Predictive Model

Use feature vectors from step #4

Evaluate & Refine Model

Perhaps return to step #2, 3, or 4



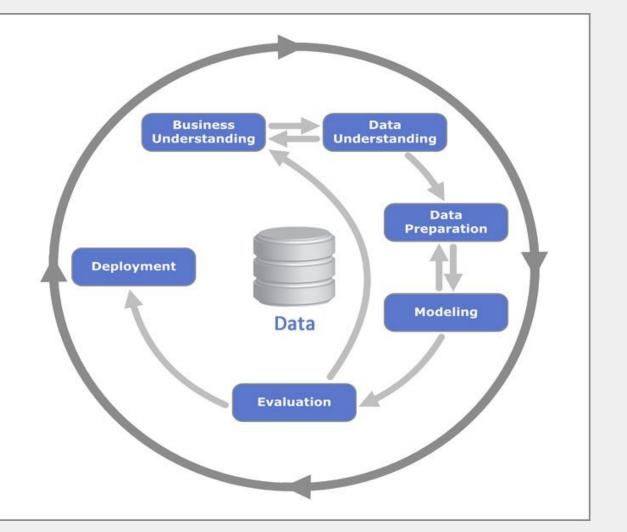
Play Around With the Data

 Split, segment, & plot the data in different ways

Identify Patterns & Extract Features

Use statistics to identify & test significant variables

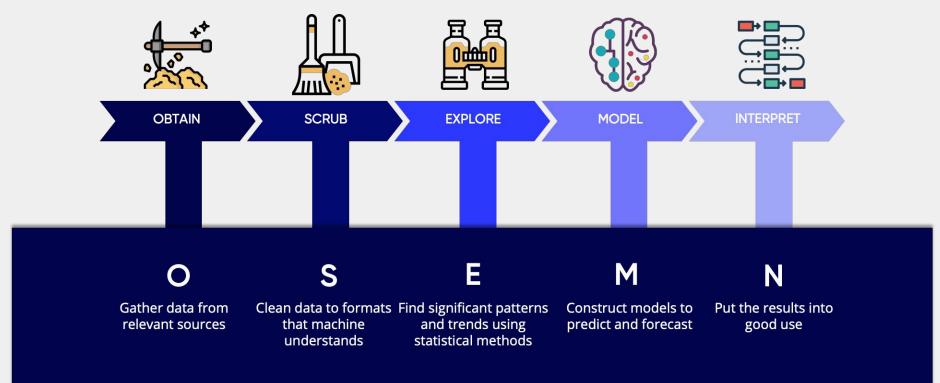
CRISP-DM Process Diagram



Source: Kenneth Jensen

Data Science Process





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